STATISTICS

W B BAILEY
AND
JOHN CUMMINGS

The National Social Science Series

Edited by Frank L. McVey, Ph.D., LL.D., President of the University of Kentucky

Now Ready: Each, Sixty Cents Net

PROPERTY AND SOCIETY. A. A. Bruce, Associate Justice Supreme Court, North Dakota, Commissioner on Uniform State Laws, etc.

WOMEN WORKERS AND SOCIETY. Annie M. MacLean, Assistant Professor of Sociology, The University of Chicago.

SOCIOLOGY. John M. Gillette, Professor of Sociology, The University of North Dakota.

THE FAMILY AND SOCIETY. John M. Gillette.

THE AMERICAN CITY. Henry C. Wright, First Deputy Commissioner Department of Public Charities, New York City.

GOVERNMENT FINANCE IN THE UNITED STATES. Carl C. Plehn, Professor of Finance, The University of California.

THE COST OF LIVING. Walter E. Clark, Professor and Head of the Department of Political Science, The College of the City of New York.

TRUSTS AND COMPETITION. John F. Crowell, Associate Editor of the *Wall Street Journal.*

MONEY. William A. Scott, Director of the Course in Commerce, and Professor of Political Economy, The University of Wisconsin.

BANKING. William A. Scott.

TAXATION. C. B. Fillebrown, President Massachusetts Single Tax League.

THE CAUSE AND CURE OF CRIME. Charles R. Henderson, late Professor of Sociology, The University of Chicago.

THE STATE AND GOVERNMENT. Jeremiah S. Young, Professor of Political Science, The University of Minnesota.

SOCIAL ENVIRONMENT. G. R. DAVIES, Assistant Professor of History and Sociology, The University of North Dakota.

THE PSYCHOLOGY OF CITIZENSHIP. ARLAND D. WEEKS, Professor of Education, North Dakota Agricultural College.

STATISTICS. WM. B. BAILEY, Professor of Practical Philanthropy, Yale University, and JOHN CUMMINGS, Expert Special Agent, Bureau of the Census.

In Preparation

WOMEN AND THE FRANCHISE. JOSEPHINE SCHAIN, Executive Secretary, Association of Neighborhood Workers, New York City.

SOCIAL INSURANCE IN THE UNITED STATES. GURDON RANSOM MILLER, Professor of Sociology and Economics, Colorado State Teachers' College.

THE MONROE DOCTRINE. A. B. HALL, Professor of Political Science, University of Wisconsin.

THE NEWSPAPER AS A SOCIAL FACTOR. ALLAN D. ALBERT, Former Editor *Minneapolis Tribune.*

THE STRUGGLE FOR LAND IN AMERICA. CHARLES W. HOLMAN, Expert of United States Industrial Commission.

MODERN PHILANTHROPY. EUGENE T. LIES, General Superintendent, Chicago United Charities.

SOCIAL AND ECONOMIC LEGISLATION. JEREMIAH S. YOUNG.

POPULATION. E. DANA DURAND, Professor of Statistics, The University of Minnesota.

COOPERATION. L. D. H. WELD, Professor of Business Administration, Yale University.

THE PUBLIC LIBRARY AS A SOCIAL FACTOR. W. D. JOHNSTON, Librarian, St. Paul Public Library.

STATISTICS

By

William B. Bailey, Ph.D.
Professor of Practical Philanthropy
in Yale University

And

John Cummings, Ph.D.
Expert Special Agent, Bureau
of the Census

CHICAGO
A. C. McCLURG & CO.
1917

Published December, 1917

W. F. HALL PRINTING COMPANY, CHICAGO

EDITOR'S PREFACE

THE value of a knowledge of statistics grows every day in the fields of business, government and social work. Most of the books devoted to its acquirement are either intricate or too extended. This book has been written with the purpose of avoiding these difficulties. Professor Bailey has had especially before him the needs of social workers for a simple book in the statistical field and with the helpful suggestions of Dr. John Cummings of the United States Department of Commerce presents a useful book that will fill the statistical gap in the National Social Science Series to the satisfaction of its readers.

F. L. M.

AUTHORS' PREFACE

THERE is an increasing number of persons in this country who require an acquaintance with the elements of statistics. Social workers must keep records and prepare reports which give an accurate and adequate picture of the year's activities. Persons in many lines of activity find it necessary to gather primary statistical facts, plan tables, tabulate the raw material, and present it to the public in such shape that it can be used for intelligent analysis and comparison. Many students desire some knowledge of the fundamentals of this subject, although they may never take advanced instruction in it.

It is to meet the needs along these lines that this small volume has been prepared. Wherever possible, mathematical formulas have been omitted. A bibliography has been added for the use of those who wish to pursue this subject further. We would express our indebtedness to Mr. Julius H. Parmelee, Statistician of the Bureau of Railway Economics, for his advice and suggestions throughout the preparation of the manuscript.

W. B. B.
J. C.

CONTENTS

Contents

STATISTICS

STATISTICS

CHAPTER I

IMPORTANCE OF STATISTICS

THE principal use of statistics is to interpret large groups or series of numbers. Pages of figures are bewildering to the average reader who is interested primarily in the lessons to be taught by these columns. He has been told that anything can be proved by figures and at the sight of a disconcerting array of them he is quite ready to believe it. Notwithstanding the discredit which is often cast upon statistics, they are being used more and more, and the number of statistical investigations which are being conducted in this country is quite appalling.

Statistics are, in fact, the language in which social conditions are accurately described and social laws accurately stated. It is the function of the statistician to write and to interpret this language which is becoming the universal language of the social sciences. Men of various occupations and interests are called upon to conduct statistical investigations or to interpret the statistical results of such investigations. A knowledge of the methods and value, of the limits and capacity of statistical inquiry

was never so essential as at the present time. Ministers are making surveys of their parishes, philanthropic organizations throughout the country are expected to keep records of their activities and to present the results of these activities in statistical statements at their annual meetings. Large corporations require in tabular or diagrammatic form summary statements of the results of past and present business activity. City, state, and national governments are turning out each year bulky volumes of statistics, and economists, sociologists, health officials, and magazine writers undertake to determine for their own guidance and to interpret for the public the meaning of this accumulating mass of data. In view of this general resort to statistics by those who have had no special training in statistical methods, it is not surprising that statistics have been frequently misinterpreted and that wrong methods have been frequently employed with the result that statistics in general have been discredited. Since, therefore, so many classes of persons are called upon to use statistics in some way, it is essential that they acquire some knowledge of the elementary principles which must be observed in every instance of statistical inquiry.

Reliability of Statistical Data

In any study which is to be based upon statistical material already available in published form, the reliability of the data available must be carefully determined.

Statistics are perhaps most frequently used when

persons consult an almanac or work of reference for desired information upon some subject. The principal danger in this use of statistics lies in giving equal weight to all figures regardless of their source and value. An examination of the sources will commonly develop the fact that some of the figures have been taken directly from official publications and can therefore be trusted, while others are little better than estimates of uncertain origin. In every case the figures should be traced to their source, to determine whether or not it is such as to inspire confidence. In some cases where the figures are found to be accurate, the number reported may be too small to warrant drawing a conclusion from them.

In every case, also, it is well to read the text accompanying a table to determine the exact meaning of the figures. It will often become apparent that two sets of figures which it was hoped to compare were not gathered under such circumstances as to render comparison possible. Thus, the statistics of the value of imports in one country may include the amount of import duties, which in another country will not be included. In some countries the number of births includes still births, while in others this is not the case. Some figures may relate to calendar years, while others may refer to fiscal years ending on March 31, June 30, or September 30. Seasonal differences in conditions may render a comparison of fiscal and calendar years extremely inaccurate. Before the attempt is made to compare figures they must be proven uniform.

There is, fortunately, an increasing tendency on the part of officials to acknowledge the fallibility of figures collected and published by them, and to endeavor to measure the probable error. Thus census figures with regard to the distribution of population according to sex are quite accurate, while statistics of the distribution of the population by single years of age are admittedly inaccurate. Official and unofficial tables frequently carry footnotes or explanatory matter to the effect that the probability of error in the figures is sufficiently great to require extreme caution in their use. It may be noted that in many lines financial and trade statistics are more accurate than social statistics. Even where a student approaches a statistical problem with an open and unbiased mind he is certain to make mistakes if he does not carefully consider the reliability of the data. The results reached by a biased writer are of course certain to be in a much greater degree erroneous. He is searching for figures which will support his contention, and wherever he finds such he is disposed to give them credence irrespective of their authority, while he may overlook or view with suspicion equally authoritative figures which do not support his argument. The conclusions of such a writer are those of an advocate, not those of an investigator.

CHAPTER II

THE FIELD OF STUDY

WHILE in the case of any given inquiry it may be found that the raw material for the study has already been collected, tabulated, and published, obviously all statistical work starts with the collection of the original data in the field, and in any statistical investigation the first point to decide is what field to cover and what information to secure.

If the study concerns merely the operations of some particular corporation, or the income and expenses of a municipality, the field is clearly limited. If, on the other hand, it be a study of the differences in the ratio between the sexes in city and country, the scope of the inquiry is less clearly defined. In most cases, however, what may be called the area of enumeration or record to be covered is fairly well determined. An organization keeps a record of its activities and the range of these activities determines the field to be covered by the records. The records gathered by municipalities, states, and the national governments, are generally coextensive with their territory.

In this connection it may be pointed out that there are some statistical investigations which, since the area to be covered is small, can be made by single

individuals or small groups of individuals, while others, which cover an extensive territory and require the filling out of a large number of schedules, may best be undertaken by some constituted authority. Thus a survey of a small rural community or an investigation of living conditions in a few congested city blocks may be made by one person, but no individual could accomplish an enumeration of the population of a large city.

The advantages of investigation on a large scale may be combined with those of intensive work, which is possible only on a small scale, by adoption of what is called the representative method of statistical inquiry. By this method, instead of studying the entire area, certain sections which are supposed to be representative of the whole are selected and carefully studied. If the selection is wisely made, a satisfactory result is obtained. The Immigration Commission adopted this method in its statistical investigation concerning the comparative fecundity of the wives of native stock and those of foreign birth or parentage in this country. It was impossible for the Commission with the resources at its disposal to cover the whole country, although the raw material was available in the Bureau of the Census. It was therefore decided to study the state of Rhode Island, the city of Cleveland and forty-eight rural counties in the state of Ohio, and the city of Minneapolis and twenty-one rural counties in the state of Minnesota. The data for these selected sections would, it was believed, reproduce in miniature the conditions which obtained in the

country as a whole, and the results of the investigation seem to have justified this belief.

Where the total number of cases in the entire field is limited, however, it is seldom justifiable to use the representative method by confining the investigation to a selection either of areas or cases, since in such instances the whole number of cases will probably be none too large to provide an adequate basis for statistical induction. In a study recently made in an American city of the girls brought before the courts during a few months, the number of cases was clearly inadequate for such induction. The total number of arrests was small, and the paper published to show the results of the investigation had little more statistical basis than a detailed study of twenty-seven cases.

A popular instance of the use of the representative method is found in the straw votes taken previous to election by newspapers. Individuals chosen at random are asked to select the candidate they favor, and from the replies received it is hoped to divine the probable result of the election. This is an example of the representative statistical method, but the results obtained from these newspaper attempts have not in the past been such as to inspire confidence in their reliability.

CHAPTER III

GATHERING THE RAW MATERIAL

SIMILAR problems confront the corporation, organization, or department preparing blanks for keeping records of their activity, and the persons responsible for the planning of an investigation or for the taking of a census. Briefly stated, it is to obtain the maximum of reliable information with the minimum number of questions. There are, however, certain general principles which should govern the preparation of the blanks or schedules.

1. Preparation of Inquiry Blanks

(*a*).—In so far as possible, the blanks should call for information which will be comparable with similar information obtained as a result of former investigations. One great value of statistics lies in the fact that it makes comparison possible, and it is frequently a waste of time and money to gather information which cannot be used for this prime purpose.

(*b*).—The questions should be clear and intelligible to the class of people to whom addressed. No excuse can justify the sending out of a blank carrying ambiguous questions to which elaborate notes are attached explaining the exact meaning of the queries. Questions to which an answer of "Yes" or

"No" can be given, generally offer but little opportunity for confusion, provided the question is clearly stated. Where, however, those replying are required to exercise a personal judgment, the difficulties in the way of securing accurate answers are much more serious. If users of intoxicants, for example, are asked to report whether they are occasional, moderate, regular, or excessive drinkers, the value of the replies will be inconsiderable, owing to the vagueness of the inquiry and to the indisposition of drinkers in any case to admit that they are intemperate. In a recent tenement house investigation the cellars were reported as "filthy," "very dirty," "dirty," "fairly clean," "clean," or "excellent." If only one investigator had been engaged in this task the problem of classification might have been solved with some measure of uniformity, but since several investigators were employed the personal equation was bound to enter. Even if this difficulty were eliminated to a considerable extent by carefully instructing the investigators, it would be difficult so to word the report as to make the public understand the exact difference between "dirty" and "very dirty," or even between "dirty" and "filthy," when applied to a tenement house cellar.

(*c*).—The value of the information to be gained from any inquiry should more than offset the trouble and expense required to gain it. It might be interesting to know a great many trivial facts, but the expense and trouble required to gather the facts would more than outweigh the benefit resulting from the investigation. It frequently hap-

pens that complicated blanks are received by organizations with the request that they be filled in and returned to some unofficial person in search of information. Oftentimes it would seem the persons making such requests can have but a vague idea of the amount of time which would be required to complete their schedules. As an illustration of such unreasonable requests, it may be noted that the charitable organizations of one state recently received from an individual who was anxious to get a bill through the legislature, requests for information in such detail that probably not one of the organizations could have furnished the information called for without keeping its entire office force steadily engaged at the task for at least a month. The impropriety and futility of such requests are apparent.

(*d*).—The questions should not be such as to excite the suspicion or resentment of the persons from whom the desired information is to be obtained. Inquiries concerning the religion, morals, personal habits, income and expenditures of an individual are likely to arouse distrust. In 1912 an investigation was made of the social condition of the inhabitants of several congested blocks in an American city. Two questions on a somewhat elaborate blank were certainly ill-advised; namely, "How many members of the family were convicted of crime during the year 1911?" "Did the family receive any private or public relief during the year 1911? State the amount." Not only did the replies to these questions possess absolutely no value,

but they served to cast discredit upon the entire investigation.

(*e*).—Wherever the information recorded on a blank is of such a nature that various items enter into the composition of a total, it is preferable to arrange these items in a column rather than side by side. In this way the addition of the figures is facilitated and it is much easier to check the totals than if the figures are to be added from left to right or right to left.

(*f*).—Where the person filling a schedule is presented with a choice of two alternative answers to a question, it should be clearly stated on the schedule whether the one selected is to be checked or underlined or the one rejected is to be crossed out. It is usually preferable to have a line drawn through the one rejected.

Where records are made on cards to be preserved for future reference, it is advisable to adopt a card of standard size. The problem of finding cabinets in which they will fit is thus much simplified. In some cases cards or sheets of different colors may be used. Thus a charity organization society could keep its record of resident cases on white sheets, and those of non-resident cases on blue. A visiting nurse association might keep its general cases on white cards and its tuberculosis cases on red cards of the same size. This reduces the liability of error from filing a card in the wrong case. The number or name used as the index in filing the card should be at the extreme top in order to facilitate the search for a particular card. Cards should be stiff

enough to stand erect, and not so thick as to make their use on the typewriter difficult.

2. *Filling Out the Schedules*

There are two methods by which the schedule or blank may be filled out. One is to send it directly to the person from whom the information is desired and ask him to fill it out. The other is to place it in the hands of an enumerator or investigator and leave to him the task of entering the information upon the schedule. Which is the better method for any particular investigation will depend upon the nature of the information to be obtained, and the number and intelligence of the people from whom the information is sought.

If a national association engaged in a certain line of work desired to obtain information concerning the activity of its local organizations, it is probable that a questionnaire would be sent to each organization with the request that the blank be filled out and returned to central headquarters. Devotion to the objects of the organization would probably lead most of the local organizations to carry out the wishes of the national association in this matter.

In the case of individuals, however, it is extremely difficult to persuade them to fill out and return a schedule unless there is some compelling power behind the request. The Department of Agriculture, for example, recently attempted to secure returns from farmers upon schedules sent out through the mails, but found after a persistent effort on the part of the Department to secure the return of these

schedules, that a large proportion of them remained in the hands of the farmers. Corporation tax and income tax schedules can, however, be handled in this way, and the statistics of railways published by the Interstate Commerce Commission are compiled from schedules filled out by the companies in compliance with a legislative requirement.

In an enumeration of the population it is more satisfactory carefully to instruct an enumerator and send him from house to house to collect the desired information. This is the census practice in this country. Over 70,000 enumerators were sent out on April 15, 1910, to record the facts concerning the population of the United States. Thirty-two questions were asked concerning every person. Most of these questions were extremely simple, but in some cases it was necessary for the enumerator to explain the meaning. It would have been practically impossible to obtain this information by leaving blanks at the houses and calling for them later on. In some families there was no one who was able to read and write English, and it is certain that the blanks would have been inaccurately filled out, ignored, mislaid, or forgotten in so large a proportion of cases that the population returns as a whole would have had no value. Where enumerators were unable to find persons at home, individual census slips were left with the request that they be filled out and ready for the enumerator when he should call for them. A vast majority of the schedules, however, were filled by the enumerators.

At the Thirteenth Census, in the case of those

schedules which carried inquiries less simple than those upon the population schedule, enumerators were advised to deliver the schedules some days in advance of the enumeration. This was done with the agricultural schedules, since it seemed advisable to give the farmer some time for the preparation of the replies to the more elaborate questions. It was found that in a large proportion of the cases, however, the farmers failed to avail themselves of the opportunity of becoming acquainted with the schedule in advance, and that it was necessary for the enumerator to spend some time in explaining the questions and entering the replies.

About the only difference between collecting the data in an investigation, survey, or census, and keeping the statistical records for an organization, concerns the time and place at which the cards are filled. In an investigation, the blanks are put in the hands of the enumerator who goes from house to house to record the replies, and the attempt is made to gather the information in the shortest possible time. The purpose is to obtain a cross-section view of society upon a certain date. On the other hand, an organization is continually keeping records of its activities, and cards are filled from day to day as cases appear in the office. In both cases the records should be so filed as to make possible future reference to any particular card. In an investigation covering a city it is perhaps sufficient to file away the cards by squares or blocks, but no such simple method of filing is sufficient in the case of records kept by an organization, since it may be

necessary from time to time in any individual case to add information to the record as the case develops. This requires frequent consultation of the records which should, therefore, be so classified and indexed as to facilitate the finding of any given record without loss of time.

Since the records of most societies are too elaborate to allow all the information to be placed upon a card, it is necessary to keep upon the card sufficient information to identify the case, and file the detailed information concerning it in another place. The cards, therefore, form little more than an index to the work of the year. The cards are usually filed in alphabetical order in cases, and each card carries the number corresponding to the number on the more elaborate record. Many organizations find it more convenient to use the visible card index rather than to file the cards alphabetically in cases or drawers. By the visible index the cards are arranged alphabetically upon metal frames so that all the addresses are visible at a glance. The cards are so made that any information contained on them in addition to the name, address, and number of the case can be easily consulted. The Rand System and the Index Visible are the most commonly used of these filing devices.

Since the larger records often contain correspondence, it is advisable to have the folders containing them about the same size as ordinary letter paper. Large envelopes are sometimes used to hold the correspondence and record sheets. More frequently a piece of heavy Manila paper is employed. This should be about eleven and a half by eighteen inches,

and so folded that the back edge will extend about half an inch above the front edge. At the top of the back edge on the projecting half-inch is plainly written the number corresponding to the number upon the index card. The folder is then placed in a file or drawer, folded edge down and number showing. This makes it possible to find any record desired in a few moments.

For convenience in making out the annual report of an organization, the cards covering the cases which have been handled since the last annual report are often kept in a separate file. At the close of the year these are put in their proper place with the cards of previous years. As an aid in determining the character of the work done by the organization, metal signals are often attached to the cards. These can be had in different shapes and colors and are easily attached to the top of the cards so as to project like little flags. Thus, if an organization engaged in charitable work wished to keep track of the principal causes of poverty in the cases with which it had dealt during the year, metal signals of different colors could be attached to the cards. Green, for example, might mean unemployment; white, sickness; black, death of the bread winner. It would then be possible to tell, by counting these signals, the prevailing causes of distress during the year. The work of counting is made easier when these signals are attached at a different point on the top of the cards according to the cause, so that the black signals are all in one line, the white in another, and the red or any other color each in a separate line.

CHAPTER IV

EDITING SCHEDULES

EDITING is a process preliminary to tabulation. It does not necessarily imply inaccuracies in the schedule returns, although inaccuracies, some of which can be corrected by the editor, will generally be discovered in the process of editing, and in some classes of schedules as, for example, in those making returns of financial statistics of corporations or municipalities, the correction of errors by editing may materially affect the results of the tabulation. Schedule editing is, nevertheless, even in the exceptional cases noted, primarily formal rather than corrective, since the schedule data are original, and are not subject to material revision where the several replies are consistent with one another, except by referring the schedule back to the enumerating agency, or by initiating a new enumeration.

The general purposes of schedule editing are to insure, in as high a degree as possible, (1) accuracy, (2) consistency, (3) uniformity, and (4) completeness in the schedule returns.

1. Accuracy

Certain replies may raise a presumption of error, and in some cases this presumption may be sufficient to warrant investigation and verification. In certain

foreign censuses, for example, when a person has been returned as over 100 years of age, an investigation has been made to determine whether the age has been correctly stated. The result has always been materially to reduce the number of reported centenarians. Schedules, or copies of schedules, collected by mail from manufacturing establishments or public service corporations or steam railways, after examination in the central office, are frequently returned to the reporting agencies for correction, or letters of inquiry covering certain points in the schedule are sent out calling for correct data.

Generally, however, the editor must accept the schedule as it is presented to him without further reference to the enumerating or reporting agencies.

When inconsistent or impossible replies have been entered upon the schedule as finally accepted by the central office, it must be edited into consistency; since the process of tabulation, which follows editing, exacts absolute consistency from each schedule. This editing for consistency may be regarded as being in a sense corrective, but it is so only in a very limited and special sense, since the scope of the editor's authority to revise replies is defined in the schedule itself. All schedule replies are equally original, and the only evidence competent to justify the revision of one reply is the evidence presented in other replies. In editing for consistency the editor makes such changes only as the schedule itself demands, and he exercises judgment only in determining which of two or more inconsistent replies shall be accepted as correct. Although in some cases

it may be impossible to determine with absolute certainty which reply is correct, generally it is true that a strong probability of correctness attaches to one reply, and there is the further possibility, in cases where no probability of correctness attaches to one reply rather than the other, of editing the inconsistent replies into the "no report" class.

It is extremely important that the editor should understand and observe strictly the limits upon his authority to make changes in the schedule, and it should perhaps be noted as a minor detail, first, that the editor should never make any erasures on the schedule which will obliterate the original return, and, secondly, that all revisions should be made in a distinctive ink, so that the work of the editor will always be perfectly apparent, since the work of the editor itself may be subject to revision and should in any case be perfectly distinguishable upon the schedule.

Errors subject to editorial correction in returns of financial or accounting statistics arise chiefly from misunderstandings on the part of those filling out the schedule, or from failure to make correct classifications of returns of income and expenditure in constructing balance sheets and in making up financial statements. Different practices of accounting in different concerns and in different municipalities must be reconciled so far as possible by editing. In order to avoid this difficulty the Interstate Commerce Commission has found it necessary to impose upon railroad and other corporations subject to its jurisdiction, uniform systems of accounting, pre-

scribing in detail the accounts that shall be kept, and defining precisely all items that shall enter into the capital accounts and into the income accounts. These orders of the Commission, which have been elaborated and promulgated from time to time during the past two decades, have been absolutely essential as a means of bringing in to the Commission in the annual reports from the railroad offices, data which was susceptible of tabulation. Prior to this action on the part of the Federal Commission, the various state railroad commissions had published the reports of the railroads, practically in the form in which they were made up in the several railroad offices, and these reports were so various in character that compilations of value could not be made from them. Where uniform systems of accounting have not been imposed upon corporations, schedule returns of financial data may require considerable editing.

2. *Consistency*

In editing for consistency, the first step is to determine upon a method of procedure to be followed in examining each schedule. Efficient and complete editing involves the systematic examination of all related replies in a predetermined order of examination. This sort of editing is, of course, impossible where the replies are absolutely unrelated to one another, and it is impossible as between unrelated inquiries on any schedule. It is, for example, impossible on a population schedule to check the age return against the sex return, or to check the return

of nativity or of country of birth against the return of marital condition. But many inquiries are more or less interrelated, and in such cases the reply to one inquiry determines within certain limits the replies to other inquiries. Marital condition, for example; may carry certain implication as to age, since practically all married, widowed, or divorced persons are fifteen years of age or older. A native obviously cannot have been born in a foreign country — although children born of American citizens living abroad have been classified as natives of the United States in order to avoid too great detail of tabulation.

Totals which are inconsistent with constituent items shown may be entered upon a schedule, as in the case of detail of income and expenditure which does not check up with the statement of total income and expenditure; or of detail regarding individuals in a family where the total number in the family, as stated, does not correspond with the number of individuals for which returns are made; or where a family budget is incorrectly totaled and balanced.

Generally inconsistencies are evidence of carelessness on the part of the enumerator, or of misunderstanding or ignorance on the part of the person filling out the schedule.

In some cases the inconsistency is not absolute, but is of such a nature as to make the return highly improbable. The return of certain gainful occupations in the case of women and young children, for example, while it may be highly improbable, may be nevertheless within the range of possibility. It

is highly improbable, but not impossible, that a child under fourteen years of age is or has been married. Generally, if the return is within the range of reasonable possibility, it must be accepted as correct, unless it can be corrected by some other related reply. The return that a person was the head of a family, and was employed in some gainful occupation, together with other detail on the schedule, might in some cases justify editing an inconsistent age return as "age unknown" on the strong probability that an error had been made in recording the age, possibly by omitting one figure in writing the age, as in recording a person of the age twenty years, as of the age two years.

Inconsistencies are not always apparent upon examination of individual schedules. Replies, which upon examination of individual schedules appear merely in some degree exceptional or somewhat improbable, may develop a high degree of improbability in the process of tabulation. One instance of this sort may be cited. At the census of 1900, it was found upon tabulating the returns that the number of Negroes returned as "unable to speak English" was so large as to be highly improbable. This return could not be edited out of the schedules, because it was entirely possible that any given Negro might be unable to speak English, but it was exceedingly improbable that the number unable to speak English should be so great as developed upon tabulation of the returns. Upon examination of the schedule used at this census, the probable explanation of the erroneous returns became apparent. In

contiguous columns the schedule called for answers to the inquiries as to the person's ability to read and to write and to speak English. In the case of whites, the usual and correct return to these inquiries necessitated writing "Yes, Yes, Yes," and in some cases it was "No, No, No." In the case of many illiterate Negroes, the enumerators made the partially incorrect return "No, No, No," instead of the correct return "No, No, Yes." In consequence of this accidental arrangement of columns on the schedule, the tabulation relating to ability to speak English for the Negro element had to be abandoned. At the Thirteenth Census the columns of the population schedule were rearranged, and much more accurate returns were secured to this inquiry.

In the construction of schedules it is sometimes advisable to introduce overlapping, or even duplicating inquiries, in order to provide checks for important inquiries, where the chance of error is considerable, as in the case where the inquiry calling for age is duplicated by an inquiry calling for date of birth. Inconsistent replies to such inquiries must be edited out by examination of other replies, or by an arbitrary selection of one reply as being correct. This procedure is, however, seldom justifiable, since the disadvantages of complicating the schedule more than offset any gain in accuracy in the case of individual schedules.

3. *Uniformity*

Editing for uniformity is required where replies, in themselves correct, are variously stated. Editing

of occupational returns is largely of this character. A given occupation may be designated variously in different sections of the country, or it may be variously returned from each section of the country. The return may, of course, be vague and indeterminate, as where a person is returned as a "clerk" or a "mechanic" or an "engineer" or an "artist" or an "operative."

In every case it is necessary to determine upon occupational designations which will consistently group the returns for tabulation. Moreover, since the number of occupational employments returned in any extensive inquiry may amount to several thousand—at the Thirteenth Census some 9,000 different employments were distinguished—and since many of these employments are each of them common to many different industries, and since occupational returns are frequently tabulated by industry as well as by occupation, some scheme of arbitrary symbols must generally be devised for editing the occupational returns into uniformity for tabulation. Commonly, the industry and the employment returned are designated by a simple combination of letters and figures, new symbols being assigned to each new employment discovered in the process of editing. The tabulation is then made mechanically from the symbols which have been edited on the schedules, in any combination that seems advisable when the editing has been completed. After tabulation the occupational designation is substituted for the symbol.

A minor instance of editing for uniformity is

found in the rounding out of numbers to be stated in hundreds or thousands, instead of units, or in full units instead of in fractions of a unit. This is done where the character of the data does not warrant a statement varying by small units, or fractions.

4. *Completeness*

Editing for completeness also is formal rather than corrective. This sort of editing may consist either in entering upon the schedule derivative data, or in entering replies to inquiries which have not been answered. Not infrequently, especially in schedules calling for financial data, percentages or other derived figures are required for tabulation which are not specifically called for in the schedule. These must be computed in the statistical office and edited on the schedule. On the other hand, replies called for by the schedule may be omitted, and these must be supplied, since for purposes of tabulation a definite reply must be entered on the schedule for every inquiry calling for a reply. Where no specific reply is indicated by other data on the schedule, the reply edited in must be "no report," "unknown," or some similar entry.

CHAPTER V

TABULATION

WHILE tabulation is a final process in the compilation of statistical data, it controls, or should control, the whole procedure in any statistical inquiry.

1. Scheme of Tabulation

The general scheme of tabulation is necessarily implied in the schedule, in so far, at least, as regards the primal or basic tabulations. The schedule should be formulated with reference to these tabulations and should be constructed to produce precisely the data called for in the scheme of tabulation. No questions should be entered upon the schedule which are not embraced in the scheme of tabulation, and obviously none should be omitted which the scheme of tabulation embraces.

While, therefore, tabulation is a final process, the formulation of the scheme of tabulation should be the initial process, preceding even the formulation of the schedule, which should be determined by the character of tables to be produced.

Failure to observe this fundamental principle in statistical practice, perhaps more than any other characteristic, distinguishes the work of the amateur from that of the expert, the work of the

untrained social investigator from that of the experienced, scientific statistician. The amateur investigator constructs his schedule to cover his interests, frequently with disregard of the specific requirements of tabulation. The trained statistician constructs his schedule to cover a formulated scheme of tabulation, and for him each inquiry represents a defined numerical aggregate, a column or a line of figures, a table or a series of tables.

It should not be inferred from this that the statistician may not professionally entertain social interests, or that social interests do not underlie his numerical aggregates. Such interests, in fact, inspire all scientific statistical inquiry; but given any social interest with reference to which statistical inquiry is to be undertaken, the statistician proceeds to formulate in detail the statistical statements in which the results of the inquiry must be presented. Of such statements in the case of any inquiry usually a very considerable number may be devised, more or less pertinent and significant but differing in some degree from one another, and calling for specific forms of schedule inquiries. Among these a selection must be made prior to the formulation of the schedule, if the schedule is to be perfectly conformed to the special requirements of tabulation. If it is a question of family income, for example, tabulations may be devised classifying incomes by amount, size of family, nationality or race, number of breadwinners, age and sex of breadwinners and of other members of the family, city districts, occupation and industry, number of rooms in dwelling,

average income per week or month, fluctuation in actual income per week during the period covered, constancy of employment of breadwinners, and by many other characteristics of families, occupations, breadwinners, dwellings, and communities. As a basis of tabulation, it is not sufficient that accurate data shall be obtained relating to any or all of these several characteristics, since the data may be entirely accurate and comprehensive and be nevertheless entirely incapable of tabulation. The classifications to be made in the tabulations determine specifically what inquiries shall be entered upon the schedule, and the precise terms in which those inquiries shall be expressed; whether, for example, income during a specified period shall be stated by weekly, monthly, or annual rates or amounts or averages. In a word, where the procedure is in accordance with correct statistical practice the scheme of tabulation determines specifically the precise character of the schedule inquiries. Tabulation is in every instance the final test of value for each inquiry upon the schedule, those inquiries only being justified which in tabulation produce numerical aggregates of significance.

In official statistical work, where inquiries covering specific fields are recurrent at regular intervals, the process of improving the character of the statistical compilations is largely one of refining the schedule so that it will produce precisely the data required for the compilation of tables of which every detail has been defined by past practice and tradition. Under such conditions, to gather more

or less data than is required for the established scheme of tabulation is a futile expenditure of public money, and in any statistical inquiry, official or unofficial, from the point of view of statistical expertness it is a serious blunder to enter the field of inquiry with a diffuse and over elaborated schedule, which, in proportion as it exceeds the requirements of tabulation increases needlessly the labor and cost of securing the data required, and may even be an occasion of inaccuracy in the essential returns.

An illustration of the controlling influence of tabulation over the whole process of statistical compilation is found in the development of the statistical work of the Interstate Commerce Commission. The early reports rendered annually by the railway companies to the Commission in accordance with the law, were accurate accounting statements covering the financial condition and operations of the companies. But these statements represented diverse systems of accounting, and the units of statistical compilation in them were not uniform in character. Each company put its own interpretation upon such terms as trackage, tonnage, maintenance expenditure, capital investment, betterments, depreciation charge, net and gross earning, and income. These various accounting systems and interpretations of terms largely invalidated any compilations made of the data, and to meet the requirements of statistical tabulation the Commission imposed upon the railway companies a uniform system of accounting, and promulgated its own definitions of terms descriptive of physical plant, operation, investment, income, ex-

penditure, and earnings. The elaboration of these systems of uniform accounts and reports to provide for definite tabulations occupied experts for more than a decade.

In this connection, the practice of the Bureau of the Census also may be instanced as conforming to the general principle that the schedule must develop out of the scheme of tabulation. At each decennial census Congress imposes upon the Bureau of the Census certain inquiries expressed in general terms. The problem confronting the Bureau on these occasions is not simply to devise a schedule which will produce the data indicated in the general provisions of the census act, but to devise a schedule which will meet the requirements of an elaborate scheme of tabulation established in past census reports. Such terms as race, nativity, family, illiteracy, school attendance, marital condition, improved and unimproved farm acreage, tenancy and home ownership, must be specifically defined not in legal or general terms, but in terms of tabulation units. The nativity tabulations, for example, are represented on the schedule by the inquiries calling for state or country of birth of each person enumerated, and of the father and mother of each person enumerated. The detail called for is determined by the scheme of tabulation, and under a different scheme of tabulation for the nativity data the form of inquiry would necessarily be different, calling for greater or less detail according as more or less elaborate tabulations were contemplated.

In any field of statistical inquiry in which a gov-

ernment bureau or a private organization undertakes at regular intervals the collection of data, or maintains a service for the continuous recording of data, the scheme of tabulation should be definitely established and the schedules of inquiry or of record should, after an initial period of experience, be perfectly adapted to the special requirements of that established scheme. Where, however, an inquiry is initiated in a new field, and in cases where the investigation is unique and not continuous or recurrent, it will seldom be possible to forecast the results of the inquiry with such a degree of precision as will enable the investigator to determine upon all the details of tabulation in advance of the inquiry and perfectly adapt his schedule to bring in precisely the data required. In such cases, experimentation may be entirely justifiable and inquiries may be included in the schedule in the hope that data may be secured sufficiently accurate and complete to warrant tabulation. This hope may be disappointed, and it may be found when the schedules have been examined and edited, that the data obtained under certain inquiries are valueless for tabulation. In such cases the data must be discarded, even though there be no technical difficulties in the way of tabulation.

It is, in fact, generally, if not always, quite possible mechanically to tabulate data which are essentially valueless and, if tabulated, even mendacious. There are undoubtedly cases where the futility of an inquiry becomes apparent only during the progress of the investigation, or even after the sched-

ules have been edited. In all statistical inquiries *de novo,* especially, and in all unique, as distinguished from recurrent or continuous investigations, the scheme of tabulation should, therefore, be carefully reconsidered after the schedules have been edited, and the character of the data is finally determined. The degree of incompleteness, inaccuracy, and vagueness characterizing certain replies may necessitate a material modification of the original scheme of tabulation.

The difficulty of forecasting the results of an inquiry and the possibility, or even probability, that the scheme of tabulation as determined upon in advance will require revision when the data are in hand, does not, however, relieve the investigator of the responsibility for devising some scheme of tabulation in advance. Unless there is at least a fair prospect of securing data for tabulation under any given inquiry, it should not be entered upon the schedule, and the justification of every schedule inquiry must be found in the provisional scheme of tabulation.

Unless such a scheme is perfected, and the schedule devised with reference to it, the data will almost inevitably prove to be seriously defective for purposes of tabulation, since tabulation, as has been noted, requires not simply accuracy and completeness in the returns, but a very specific sort of accuracy. Returns of wages may be accurate, and yet be untabulable, and they will almost certainly be defective for any sort of tabulation whatever, unless they have been secured under some definite scheme

of tabulation. This scheme will necessarily determine precisely in what terms wages shall be reported; whether in rates, or in amounts earned, per hour, day, week, month, or year, for full time or for time worked, including or excluding overtime work, including or excluding board or lodging or other items. Statistical tables are essentially specific in their meaning, and they require data that are uniformly specific in the same kind and degree.

The provisional scheme of tabulation necessarily embraces all possible replies to each inquiry entered on the schedule, and in initiating any statistical inquiry the number of possible replies to each inquiry, and the nature of the replies, should be carefully considered as units of tabulation.

Every inquiry implies at least two replies, and in many instances the number of possible replies is limited to two. All inquiries calling for a "Yes" or "No" answer are of this character: such, for example, as the inquiries on the population schedule relating to illiteracy (Can the person write?); school attendance (Has the person attended school?); and unemployment (Was the person at work on April 15?). Other inquiries which do not call specifically for a "Yes" or "No" answer may nevertheless admit of only two replies, and might in all such cases be thrown into a form calling for a "Yes" or "No" reply. Such an inquiry is that relating to the blind and to the deaf and dumb. In the case of these inquiries blind persons are returned as blind, and deaf and dumb persons as such, no entry being made in these schedule columns

for persons who are not blind and are not deaf and dumb. In these cases blanks on the schedule indicate in one case that the person was not blind, and in the other not deaf or dumb. Such inquiries, it may be remarked, constitute exceptions to the statement above that all inquiries call for at least two replies, unless the blank is regarded as being constructively a reply. The blind inquiry might take the form, "Was the person blind in both eyes?" If it were in this form, an entry on the schedule "Yes" or "No" would be called for in the case not only of blind persons as being blind, but of all persons enumerated as either blind or not blind. Obviously this would add materially to the labor of enumerators. Other inquiries admitting of only two replies are those relating to home ownership (Is the home owned or rented? Free or mortgaged? Farm or house?); citizenship (Was the person naturalized or alien?), and to sex.

In the case of certain inquiries admitting of only two replies, a blank space or omission of any reply is constructively a third reply. Such, for example, is the inquiry relating to Union and Confederate veterans, where an entry must indicate a survivor of the Union or of the Confederate army or navy, and the blank indicates that the person enumerated was not a veteran. This inquiry admitting of only two replies, distinguishes three classes of persons —Union veterans, Confederate veterans, and persons who are not veterans.

In the case of one occupational inquiry (Is the person an employer, an employee, or working on

own account?), the number of possible replies is three. In the case of the marital condition inquiry it is four — single, married, widowed, or divorced. In the case of the age inquiry, calling for a statement of age by single years, with detail by months for children under two years of age, the number of possible replies exceeds one hundred. In the case of the nativity inquiry, calling for a statement of the state, territory, or country of birth, it equals the number of geographical areas defined as states, territories, or countries; and in the case of the inquiry calling for a specific statement of occupation, the number of possible replies amounts to several thousand. The occupational designations developed in the primary editing of the Thirteenth Census occupation returns were reduced by several thousand in tabulations by combinations of similar employments. Although the number of replies which will be made to other inquiries, such as those relating to duration of present marriage, number of children born and surviving at date of enumeration, year of immigration to the United States, and mother tongue, cannot be precisely determined in advance, maximum and minimum limits can be designated.

In the case of other inquiries differing essentially from those of the Census population schedule — such as those calling for a statement of amount of capital investment, income and expenditure of corporations, number of employees, tonnage of different commodities moved by carriers, quality and value of imports or exports, number of farm ani-

mals of different classes, acreage of improved land in specified crops, number of pupils enrolled in public schools, revenue and expenditure of municipalities — the number of possible replies is indefinitely great, and may equal the number of individual schedules.

It will be noted that in the case of certain inquiries, even of some admitting of a large number of replies, all admissible replies are known and can be written down before the investigation is initiated. In the case of other inquiries, including those under which are reported aggregates of number or amount, although the exact replies cannot be determined in advance, the units enumerated are entirely simple in character, and constitute in the aggregate simple categories of data. In one class of inquiries each reply enters into tabulation as a unit, as representing an illiterate or a literate person, a child attending or not attending school, a person able or not able to speak English or some other language, a person employed or unemployed, a male or a female, a single, married, widowed, or divorced person, a person of a specified age, a person born in a specified state, territory, or country, a person working at a specified occupation, a blind or deaf and dumb person. In the second class of inquiries, the unit of number, value, or amount is an undifferentiated simple unit, each schedule reply representing not a single unit, but a statistical aggregate of such units.

Thus, two general classes of statistical inquiries may be defined, as indicated above, namely, (1)

those in which single units are returned in categories set up in the inquiry itself, and (2) those in which simple units are returned in aggregates.

As regards inquiries which set up categories in the schedule returns, it is to be presumed that these categories have, except in cases where they are exceedingly numerous, been definitely determined upon in advance, and that they are clearly implied in the schedule inquiry itself. The census inquiry relating to sex clearly implies either "male" or "female," as the only reply admissible, and the inquiry relating to marital condition clearly implies as the only reply admissible in case of known marital condition, either "single," "married," "widowed," or "divorced." For such inquiries the tabulation classes are determined, since these classes must correspond with the categories of the schedule data. The primary sex tables will distinguish two classes, the marital condition tables four, the age tables one hundred or more, the nativity tables as many classes as there are nativity areas, and the occupational tables as many as there are occupations reported. In each case the primary tabulation process is a process of finding aggregates for the several categories, the number in each class, by summing up the units reported on the schedules.

As regards the second class of inquiries, those returning aggregates of simple units, each inquiry represents not several, but one single class in the tabulation scheme—passenger train mileage, acreage in cotton, dairy cows, capital invested, value of product, cost of materials, bushels of corn, or

other aggregate specified by the inquiry. The primary tabulation of such returns may be a simple summing up, in this case a summing up not of units by categories, but of schedule aggregates in a single category. In many instances, however, the tabulation may properly be involved by distinctions imposed upon the data arbitrarily. The classification of farm acreage, for example, and of all farm data, may be subjected to an arbitrary distinction separating farms into classes according to acreage reported by individual farms, aggregates being obtained for small farms separately and for large farms of specified acreage.

The value of tabulations will be largely affected by the character of these arbitrary distinctions, and only very general principles of guidance to be observed in making these distinctions can be indicated.

It is of fundamental importance in devising any scheme of classes, that the distinctions shall be such as will produce aggregates that are socially significant and only such aggregates. The tabulation should not be involved by merely mechanical distinctions which by multiplying classes may obscure rather than develop the significance of the data, and should not, on the other hand, be so simplified and reduced as to blanket or ignore distinctions of importance. It is sometimes the case in official and in other statistics that the mechanics of tabulation intrude upon the data. Under a mechanical differentiation, age classes may be formed by decimal or quinquennial periods only, and significant cleavages in the age or structure of the population ignored.

Obviously the age classification should recognize such important lines of division as the age of legal majority, the militia age, the child-bearing age, the age of compulsory school attendance, and the age limits specified in child-labor laws. Much greater detail of age is required for certain periods of life — such as early infancy and generally the period of adolescence — than is required for other periods. A merely mechanical tabulation by age periods of uniform deviation does not satisfy these requirements. Such a grouping will be, over considerable life periods, too detailed for convenient utilization, and over other periods too broad to develop significant totals, although in the primary-age tables, of course, groupings by uniform periods of one, five, and ten years, are essential as well as the irregular groupings indicated above.

A second general principle of tabulation classification may be noted which is frequently of prime importance. In every instance, so far as possible, comparability of the date with corresponding data in other similar inquiries, past and present, should be preserved. In some cases an obvious improvement in classification may wisely be avoided where it involves impairment of comparability, since the actual condition at any given time or in any given locality may have significance largely by relation to conditions which have obtained in the past, or obtain at present in other communities, a relation dependent upon uniform classifications.

A third principle involves the extent of the inquiry. The tabulation should not be too detailed

for the data. That amount of detail which is warranted by an inquiry embracing the country as a whole may be altogether insignificant if shown in an inquiry restricted to a small area. The detail of tabulation must be, to a considerable extent, regulated by the number of schedules in hand, so as to avoid tables in which the data are spread out too thin to carry weight on the one hand, and, on the other, tables in which the data are massed in undifferentiated aggregates.

Finally, the detail of tabulation must represent fairly the significance of the data and must not impose a significance not inherent in the data. A classification of cases by race or nativity, for example, implies significance in the racial or nativity factor, and may be misleading where these factors are not, in fact, important or where the several racial elements are not fairly represented by the schedules.

In general, it is in making up his scheme of tabulation that the statistician encounters his most perplexing problems. This scheme must embrace the new data in significant categories for the special inquiry in hand. It must frequently embrace more or less defective classifications established by official practice in the past. It must avoid over-refinement and over-tabulation, and at the same time provide that degree of elaboration which is essential. It must observe the limitations imposed by the mechanical devices for tabulation where the data are to be run off on machines, and it must avoid distinctions which greatly increase the labor and expense of

tabulation without adding proportionally to the value of the results. In devising his scheme of tabulation, the statistician must, by the exercise of what may be called statistical intuition or imagination, detect the significance of the unrefined data; he must hold clearly in mind the specific purposes of the inquiry, all the processes of tabulation, and the precise formal results which those processes under the devised scheme of tabulation will yield. Even the capacity of the printed page to carry tables of different dimensions in the make-up may become an extremely important factor, and generally it cannot be entirely disregarded. At no point in his investigation should the statistician proceed with more caution and circumspection, since once the tabulation is in process under the scheme, revision of the scheme is practically impossible, and all subsequent developments are necessarily finally conditioned and limited. Distinctions which have been omitted cannot be introduced, and those which have been made frequently cannot be eliminated without a very considerable expenditure of labor and time.

Some technical details of statistical practice in tabulation under any scheme of classification may be briefly noted, and it should be remarked that tabulations are frequently rendered valueless by carelessness in respect to these details.

After the schedules have been edited, they should be examined to determine whether the replies to all of the questions are sufficiently complete to warrant tabulation. It may appear that certain questions have been answered in so small a proportion

of cases as to preclude the possibility of gaining any dependable information by compilation of the data.

As regards the inquiries to be tabulated, the first step in planning the tabulation is the determination of the number of different replies made to each question. This will govern the number of spaces which must be allotted to each question in the classification of the replies. A separate space should be reserved for all replies which are most frequent, and, as far as possible, for those of minor frequency. It is always possible to combine one or more minor groups for purposes of printing after the detailed tabulation has been completed, but it is never possible to add to the detail of a table without going over all of the schedules again. It is advisable to show every set of facts in detail at least once, but where these are combined with other facts some of the detail may be omitted. Thus in a census of population it is well to tabulate the ages of the population by single years, and to publish these detailed figures with their distribution by sex and nativity. Such tables expose a high degree of error in the return of age, but they enable investigators to make special combinations of ages not shown in the census tables. When combined with factors of marital condition, illiteracy, or occupation, it is sufficient to show the age distribution by suitable age periods. In a table of two dimensions the factor of primary importance should be shown in more detail than the one which is secondary.

In most tabulation it is necessary to reserve one

space for "all other," and a second for "unknown." Even when provision has been made for classification in considerable detail, it frequently happens that cases appear which do not seem to fit into any category. Such may be thrown into the group "all other." While figures may be given in considerable detail in one table, it is not always necessary to repeat this detail in all tables. Limitation of space in printing may make it necessary to unite certain groups, or even to include some of the less important groups with "all other."

It is rarely the case in statistical investigations that all of the questions upon all of the schedules have been answered. Wherever a question has been unanswered, or the reply is too indefinite to permit satisfactory tabulation, it must be checked off in the "unknown" column. It is better policy frankly to admit, where such is the case, that upon certain subjects the replies possessed limited classificatory value, than to endeavor to force them into groups to which they may or may not properly belong, since there is no surer way to discredit an investigation than to claim for the results a degree of precision and accuracy which does not in fact characterize the data.

Where it is difficult to determine in which particular group a case falls, it is advisable to formulate rules to cover doubtful cases and then state in a footnote to the table or in the accompanying text the method followed. Thus a charitable society may desire to classify according to the causes of poverty the cases with which it has dealt during a

given year. In one case there may have been a call for assistance because the lazy, drunken husband would not support the family. Should the primary cause designated be unemployment or shiftlessness or drink? Each case must be studied carefully to determine the cause, and a uniform method of classification must be followed in all doubtful cases. Where several persons are engaged in the process of tabulation it is necessary either to instruct them carefully in the method to be pursued in tabulating doubtful cases, or to refer all perplexing questions to one person for decision. This is the only way to secure uniformity. In every case the public should be allowed behind the scenes, that they may appreciate the difficulties of the classification and the methods by which they have been met.

2. *Hand Tabulation*

The data for tabulation relating to individual cases may be upon separate blanks, each blank representing an individual case, or upon sheets or schedules each of which represents a number of individual cases. The method of tabulation to be followed is, however, practically identical whether each blank or schedule represents one or several individuals.

Whether the tabulation shall be made by hand or with the aid of mechanical devices depends upon the number of cases. In the ordinary investigation made by an individual or small group of individuals, and in such statistical work as is involved in the preparation of annual reports by societies, the num-

ber of cases is seldom so large as to preclude tabulation by hand, and this method is generally employed in one-time investigations, such as local surveys of social conditions, where investment in mechanical equipment would be unprofitable.

It is necessary at the outset to determine the combinations or classes to be shown, since great economies of labor and much more significant results are achieved where the data are taken off systematically according to a complete scheme of tabulation, rather than by single inquiries independently of one another. It is not economical, for example, to go through all of the schedules in order to distribute the individuals according to one variable after another, since with little additional labor two or three different factors may be taken off in combination in one handling of the schedules. Thus a hospital might desire to show the sex, age, and cause of death of the patients who died in the institution during a year. These three factors might be tabulated separately by going over the schedules three times or by going over the schedules once and recording the replies upon three separate sheets. By such a procedure the distribution by sex, by age, and by cause of death would be determined, but each distribution would be independent of the others. The distribution by sex and age of those dying from the several specified causes would not be determined, and this might well be the most important fact to be brought out.

To determine this cross distribution the tabulation might be made as follows. A large sheet of

paper might be ruled into rectangles by horizontal and vertical lines, the horizontal lines being sufficiently numerous to provide a space for each cause of death, and the vertical lines sufficiently numerous to provide a space for each age group which it is desirable to show separately. Causes of death would be written in the spaces along the left margin and the age groups in the spaces along the top of the sheet. Each rectangle would represent a specific cause of death designated in the stub and a specific age period designated in the box, and each case would be checked off in the proper space according to age and cause of death. The additional distribution by sex might be obtained by recording males with a blue and females with a red pencil, by distinctive checks for males and females, or by checking males at the top and females at the bottom of each rectangle. It will, in fact, generally be found entirely feasible in the ruling off of spaces to provide separate spaces for males and females.

As it is easier to count by fives and tens, it is generally preferable in checking to make four short vertical marks for four cases and a diagonal mark crossing them for the fifth case. Another method is to complete each side of a square with four cases and draw a diagonal through the square for the fifth case. Sometimes four dots are made in a row and a line drawn through these four dots for the fifth case.

When the tabulation has been completed, the addition is facilitated by counting five at a time, and it will generally be found convenient to write on the

work sheets in the several spaces with a distinctive ink or pencil the sum of the checks. In this way the skeleton table is formed on the work sheets, and the figures can then be added up and down, and from right to left, one set of totals being entered in a column against the stub, and the other set of totals being entered in a line under the box, one grand total covering the sub-totals of lines and columns. This is an easy way of checking the addition.

It is possible to tabulate by four variables at once, but this is the maximum number it is well to attempt. In the example given, it might be desirable to show whites and colored separately for each cause of death. In this case additional horizontal or vertical lines might be drawn, subdividing the rectangles for each cause of death, reserving the upper row or the left hand column for whites and the lower or right hand for colored cases. The total number of rectangles would thus be doubled, and would be doubled again if a fifth characteristic were distinguished. Obviously the liability)to error in checking, in counting, and in summing up totals increases rapidly with the multiplication of spaces.

3. *Machine Tabulation*

When the cases are so numerous that it is out of the question to do the tabulation by the ordinary method of counting, mechanical assistance is required. The first step is to transfer the information from the original sheets to cards of uniform and convenient size. This is done by a system of punches. The ordinary Hollerith card is six and

five-eighths by three and a half inches in size, and is divided into 288 squares. A punch hole in any one of these squares corresponds to some fact on the original schedule. The card is divided into fields arbitrarily, the squares of each field representing some one category of data. In the work of the Population Census of 1910, when over 91,000,000 cards were punched, a hole in one field of the card indicated sex; in another field, age; in another, the marital condition; in another, the state of birth; in another, whether able to read and write; and in other fields other characteristics. The punching was done either by a hand punch or by a punch equipped with a key-board somewhat similar to a typewriter.

As these cards were punched they were filed away, and as one corner of the cards was trimmed off, it was possible to keep them properly arranged with fields exactly superposed on one another. When the time came to add up the cases and tabulate them in various ways, the cards were run through machines.

The punched cards were first run through a verification machine which threw out all inconsistencies. Thus if an operator had made the mistake of recording a child of six years as widowed, this mistake would be detected at the start and a correct card substituted by reference back to the schedule, each card bearing identification punches which enable this sort of verification. The cards were next run through an automatic sorting machine which separated them into certain main classes. These machines handle about 300 cards a minute.

The sorted cards were then run through counting machines which counted them at the rate of 500 a minute.

The machine counting and cross tabulation in combination is effected by establishing electrical contacts through the punched holes. In the type of machine used the cards are placed over a metal frame the same size as the card. Underneath each square in the card is a hole in the frame containing mercury. Above the card as it lies on the frame is a corresponding frame which contains a needle corresponding to each possible hole in the card. These needles are attached to springs so that in case they touch the card where there is no perforation they are pushed back, but wherever there is a hole in the card the needle passing through the hole and into the cup of mercury establishes an electrical connection. By means of wiring these connections, it is possible to report a large number of combinations.

For each combination there is a separate dial and when the cards for a certain ward or township have been run through the machine, a reading is made from the dials, and the totals in the various classificatory groups are obtained by geographical areas.

In some machines the feeding of cards is automatic. In others the cards must be placed on the frame by hand. The automatic machines are capable of handling from 250 to 400 cards a minute. Were it not for these mechanical devices, the preparation of a work like the Population volumes of

the Census with the detailed classifications would be out of the question.

Large corporations are now using cards similar to these for keeping track of their records of various kinds. Boards of health and other municipal and state agencies, also are coming to use machine tabulation more extensively. By such methods tabulation in fuller detail can be undertaken, and the work completed much more expeditiously than is possible where hand methods are employed, enabling prompter publication of the report of the year's activities.

CHAPTER VI

RATIOS

WHEN the sheets containing the results of the tabulation from the schedules according to the scheme of classification have been made up, the process of primary compilation is completed. Now comes the task of placing this information before the public in convenient form.

1. Importance of Ratios

The results of the primary compilation should generally be published in detail, since these comprise the statistical results of the investigation and constitute the basis of all analytical and derivative tables. They are indispensable for many purposes. Moreover, the public has the right to know the number of cases covered by the investigation, and precisely what were the results of each inquiry, and this information cannot be withheld without discrediting the report. As a general rule a statistical report should contain all of the data of its derived figures, so that a reader may, by performing the proper processes, obtain any figure given. This is an important check against errors in the report, but it is important chiefly because it enables the reader to make analyses which, although they are not contemplated in the report, may nevertheless

be of value. No report can undertake to be complete in the sense of presenting every analysis that is of significance, but every report should be complete in the sense of presenting the data in such detail that any analysis of significance can be made by persons interested to develop the data along special lines.

The custom is growing of publishing with the detailed figures derived tables which bring out more clearly the meaning of the primary tabulations, and these derived tables generally embrace various rates or ratios. The principal advantage of this sort of derivative figure is that it substitutes for two unwieldy numbers one simple number which embraces the significance of the two, and at the same time facilitates comparisons by speaking in terms of a common denominator.

2. *Definition*

Ratio is a general term used to cover certain figures which are derived from the numbers obtained by tabulation of data, and which are used for purposes of analysis and interpretation of statistics. Ratios, together with averages, indexes, standard units, and frequency distributions, are the language in which statistics talk. In this language the meaning of the aggregates obtained by the process of tabulation is expressed and made comprehensible. Except as interpreted through these derived figures the statistical aggregates obtained by tabulation are mute, unrelated, and incomprehensible. The population aggregate has no comprehensible meaning

except by relation to other aggregates — either of population or of some other category of data. The accumulated wealth of a country is a sum which cannot be comprehended as an absolute and independent fact. But the accumulated wealth of one period can be related to the accumulated wealth at another period, or to the population, or to the wealth of other countries, and certain intelligible conclusions developed, as, for example, that wealth is increasing, or decreasing, relatively to population, that it is increasing more or less rapidly in the present as compared with the past, or in one country as compared with another.

The number of deaths (from specific diseases), suicides, crimes, persons committed to jails or asylums, freight cars operated by railways, amount of money in circulation, gold held in reserve, crop yields in any one year, exports or imports — none of these aggregates are susceptible of intelligent comprehension stated as simple, absolute, unrelated amounts. To acquire significance the number of deaths must be related to population, and similarly of each other statistical aggregate it is true that its relations must be defined in numerical terms. In the relation of one statistical aggregate to another, the entire significance of statistical compilations is determined, and these relations are largely defined in ratios.

Since a ratio is a number expressing the numerical relationship of one aggregate to another, every statistical ratio comprehends two aggregates, (each of which has been obtained by tabulation,) and may

be written as a fraction. The ratio of seventy-five to 150, for example, is one to two, or one-half. Reduced to a percentage, this ratio becomes fifty per cent, seventy-five being one-half or fifty per cent of 150. The ratio of 150 to seventy-five is, of course, two to one, or two divided by one or 200 per cent.

In practical statistical work ratios are commonly denominated: sometimes ratios, as the ratio of males to females in the population; sometimes percentages or proportions, as the percentage or proportion foreign-born in the population; sometimes distributions, as the percentage distribution of the population by age; and sometimes rates, as the rate of mortality or of natality.

These figures are all essentially ratios in the mathematical sense of that term, the various designations being adopted by common usage in statistical work. Any of them can be expressed as a fraction or as a percentage, the percentage being, in fact, equivalent to a fraction with a denominator of 100.

Commonly, statistical ratios reduce two numbers to the number of one, per unit, per 100, per 1,000, or per 100,000 of the other. The ratio of representation in Congress, for example, reduces the number of population and of representatives to population per representative. The density ratio reduces population and square miles of area to population per square mile. The rate or ratio of population growth — the increase per cent of population — reduces the population increase, and the population at the beginning of the period, to the increase per 100 popu-

lation. The death-rate reduces the number of deaths and the population to deaths per 1,000 population. The ratio of specific mortality from certain causes, reduces the number of deaths from a specific cause, such as suicide or scarlet fever, and the population to deaths per 100,000 population.

While all statistical ratios are essentially simple mathematical statements of the sort indicated, statistical practice has established certain usages which should be generally observed. Population increase might obviously be stated and is sometimes stated as increase per unit, or per thousand, or per ten thousand, rather than as increase per cent; and the crude death-rate might be generally, as it sometimes is, stated as deaths per cent of population rather than per thousand. The more common or customary form of statement should be used in every case where usage has established any given form, unless there are special reasons for using a different form, and in such special cases a precise description of the form used should be given with the reasons for adopting it, since unusual ratios are more liable to misinterpretation than are the common ratios, and are generally inconvenient for purposes of analysis involving comparisons with accumulated data expressed in conventional terms.

It may be noted that the conversion of percentages worked to one point decimal, as percentages commonly are in statistical tables, into number per unit or per 1,000 is effected simply by moving the decimal point. A population increase of eleven and two-tenths per cent is, of course, an increase of 112

per 1,000 of population. Similarly a death-rate of fifteen per 1,000 is a death-rate of one and five-tenths per cent.

Perhaps the chief reason for using percentages generally, where other forms have not been established by usage, is found in the fact that percentages are more conventional and are, partly in consequence of their conventionality, more easily referred to in the text than other ratios. A population increase is more simply referred to as an increase per cent than as an increase per thousand population, and similarly the proportion foreign-born, the proportion living in urban communities, the proportion married, the proportion mulatto are more simply referred to as percentage foreign-born, urban, married, mulatto, than as the number in each of these classes per 1,000 population. In writing text dealing freely with ratios, the fact that the language provides a single word "percentage," indicating "number per hundred," is a matter of considerable convenience.

Where statistical usage has established some other ratio than the rate per cent, as in mortality, natality, and marriage-rates, it is for some special reason. In the case of death-rates, for example, percentage rates involve small fractions, the crude death-rate being generally, in a well conditioned population, between one and two per cent, and for specific age groups only a small fraction of one per cent. It is more convenient to state these rates as number of deaths per 1,000 population, than as fractions of one death per 100 population.

In popular writing, percentages are sometimes stated as number per 100 under the impression that this statement is more easily comprehended by the reader, or as a means of varying the text, and this practice is of course perfectly allowable.

3. Classification of Ratios

Statistical ratios may be generally classified according as they express the numerical relation either of (a) totals to totals in one category of ~~data,~~ or (b) of a part or parts to a total, or (c) of a part to a part, or (d) of a number of one category to a number of some other category.

Simple illustrations of these several classes of ratios will be found in any statistical report. The a) percentage increase of population is essentially a relation of total population at one date to total population at a subsequent date, an increase of ten per cent, for example, being equivalent to a statement that the population at the end of the period amounted to 110 per 100 at the beginning of the period. In some instances tables are prepared in which the population at the end of the period is represented as 100 and populations at earlier dates expressed as percentages on this base. The percentage b) Negro in the population at any given date is a relation of a portion of the population to the total population, the number of Negroes per 100 of total population composed of Negroes and other racial classes. A percentage distribution by sex, age, or c) racial class, is a relation in each case of the several classes composing the population to the whole popu-

c) lation. The sex ratio, or number of males per 1,000 females, is a relation of one class or part of the population to another class or part, as distinguished from the percentage male, or female, which relates the number in each sex group not to the number in the other group, but to the total population. Finally, the death-rate, birth-rate, marriage-rate, rates of criminality, of population density, wealth per capita, d) and many other rates, relate numbers of one category to numbers of another category — in the cases mentioned, a relation of deaths, births, marriages, crimes, areas, and wealth, to population. Numerous other such ratios might, of course, be cited in which population is not one term of the ratio, as, for example, the ratio of earnings to investment, or of bonded indebtedness to stock outstanding.

The specific value of these derived figures for purposes of analysis and comparison will be obvious. The absolute population increases in two successive decades, since they do not in themselves indicate whether population was increasing at a more rapid rate in one decade than the other, fail to give information relating to population that is of prime importance. Even where populations are given together with the increases, only very marked changes in the rate of growth will be apparent, and these cannot be accurately estimated by an examination of the absolute numbers, unless ratios of increase are determined.

The percentage increases for two decades, on the other hand, without any population figures whatever are immensely significant, especially in a coun-

try, such as the United States, in which, since the registration of births and deaths is very incomplete, natural increase of population must be determined by enumeration. At each decade since 1790 the population of the United States has increased, the absolute increase in each decade exceeding that of the decade preceding during the entire period of 120 years, but the rate of increase in the last half of the nineteenth century showed a marked decline, which, if continued during the present century, will inevitably reduce the absolute increases, and as regards certain classes may initiate a stationary or a declining state of population. This tendency is not at all obvious in the absolute increases. The absolute increases of the several racial elements in the population, also, have in themselves, except as they are reduced to rates, comparatively little significance as regards the future racial composition of the population, while the percentage increases for these several elements are, on the other hand, full of significance, since they express the increases in terms which facilitate comparison of the rate of growth of one population element with that of another.

Similarly, the absolute urban and rural populations at the several censuses do not indicate in terms which facilitate comparison of one year with another, the extent to which urbanization of the population has taken place, and whether the drift to the cities is, relatively to population growth, increasing or decreasing, or whether it affects one class more than another.

In the field of vital statistics, ratios of mortality,

natality, marriage, and fecundity constitute in a special degree the end and aim of statistical inquiry. The number of deaths or births or marriages is of significance only when reduced to a ratio more or less refined by sex, age, race, occupation, and character of the community lived in as urban or rural.

In a word, the fundamental characteristic of ratios is that they state conditions obtaining in any year or in several years, and changes in conditions during any given period, or succession of periods, in terms which facilitate comparison of year with year, of period with period, of class with class, of one country with another, or along some other line of interest. Without such comparisons, statistical tabulations would be in many fields of inquiry futile and valueless.

Some of the simpler ratios are considered in the following sections.

4. Ratio of Increase

Comparison of population or amount of accumulated wealth, or of crop yield in any one year, or of any number or amount pertaining to any date or period with the corresponding number or amount at a subsequent date or for a subsequent period, is commonly facilitated by the introduction of ratios of increase into tables showing absolute numbers or amounts. These ratios express for each period of the table the relation of the absolute increase to the absolute number or amount. By reducing increases in the several periods to common terms, usually to terms of increase per cent, the relative increase of

one period can be compared with the relative increase in any other period, and the relative increase of any one number or quantity during any given period with the relative increase of any other number or quantity.

Ratios of increase are, perhaps, the most common form of statistical ratios, and partly because they are a simple and obvious means of indicating relative change in numerical aggregates, they are sometimes employed injudiciously, where other ratios, or indexes or averages would be more significant. Nevertheless ratios of increase generally have some value even where they are not indicated as the most significant derivatives to be determined.

As a general principle it may be stated that ratios of increase are indicated as significant, first, in cases where the increase bears a significant relationship to the base upon which the increase is figured, and especially in cases where, as is true of population growth, the increase is what may be termed an organic, as distinct from an accidental increase of the base itself; and secondly, in cases where the increase, although in itself accidental, is to be related to other consequent or concomitant increases.

A ratio of increase, formally at least, implies some such relationship either of the increase to the base, some relationship that is not entirely accidental, or of one increase to another; and where no such relationship obtains, the ratios are to that extent insignificant or even misleading. The percentage increase during any given decade of the acreage of improved farm land in different states, for ex-

ample, is in itself of comparatively little significance, because the increase in such acreage in any state is as regards the area improved at the beginning of the decade entirely independent and accidental. The percentage increase may be small in a state showing a large absolute extension of improved farm area, and may be very high in a state showing a small absolute extension, since the absolute increase is not in any way determined by the acreage improved at the beginning of the period. This, nevertheless, is one of the two factors determining the percentage increase. One of many conditions determining the extension of improved farm area is availability of land for improvement, and a relation of the increase in improved acreage in any decade to acreage unimproved and available would have in some respects more significance than a relation of the increase to acreage already improved. A percentage increase of improved acreage obviously does not indicate the rate at which the unimproved land is being taken up, or the period of exhaustion of land available for improvement—which are facts of considerable interest. For the country as a whole, however, a percentage increase of improved farm acreage has an obvious significance when related to the percentage increase of population, since in the country as a whole the relative increase of improved acreage and of population is an important factor determining the supply of food and clothing. Much less significance attaches to the relative increase of improved acreage and of population in the several states, because the crops of certain states are to a

greater or less extent consumed by the population in other states. In the several states, however, percentage increases of improved acreage may have significance when related to percentage increases of crop yields in these states, or of aggregate value of farm products.

Generally, percentage increases of aggregates of which the increase is not organic are in themselves naturally barren of significance, and require a relation to other concomitant or consequent increases to develop significance. Organic increases, on the other hand, such as that of population, are naturally significant in themselves as measures of organic development or growth.

In every case, the base and the increase should be entirely homogeneous. Frequently a slight change in the form of a schedule inquiry so changes the character of the data that no true increases can be determined, and such changes should be avoided unless they are absolutely necessary. Ratios of increase imply perfect comparability of data, and since the significance of the data is largely expressed in such ratios, the possibility of determining them constitutes generally a prime motive for preserving comparability so far as this can be done without seriously impairing the character of the data.

5. *Distributions of Aggregates*

As in the case of increases, distributions of aggregates are commonly made on a percentage basis, and are relations of parts to the whole which these parts constitute, expressed in terms of 100 units

of the whole. The employment of percentages is, however, purely a matter of usage which is not necessarily observed in every case, and does not in any case affect the principles in accordance with which distributions should be made.

These principles are quite obvious and simple. A distribution implies a degree of homogeneity in the total distributed, and at the same time a differentiation of parts or classes within the total.

An illustration of such a total is found in the aggregate population of a country, which is a clearly defined statistical aggregate, composed of many clearly differentiated elements. Each individual comprised in the aggregate population of a country possesses certain fundamental characteristics such as those of sex, age, nativity, race, and marital condition, any of which may serve as a basis of distribution. The aggregate of farm animals returned on the agricultural schedule, on the other hand, is an aggregate which does not possess a sufficient degree of homogeneity to justify a distribution by classes of animals, showing, for example, what percentage of the total number of farm animals are horses, mules, asses, cows, goats, swine, sheep, and poultry. The aggregate in this case is purely formal and as a basis of distribution fictitious. No important significance attaches to its numerical composition, since the classes of which it is composed are differentiated to a degree which constitutes them in themselves independent and unrelated aggregates.

The differentiation of classes in other aggregates developed in statistical compilation may, on the

other hand, not be sufficiently well defined to justify a distribution either absolute or relative. The classification of the Negro population by color, as black, mulatto, quadroon, and octoroon, for example, implies a differentiation which does not in fact obtain in the Negro population since this population comprises individuals of mixed blood in every degree, from the black with an imperceptible trace of white blood to the white with an imperceptible trace of Negro blood. In such cases, however, the question of making a percentage distribution is secondary. If the data are sufficiently accurate to warrant tabulation by the classes indicated, a percentage distribution by these classes will commonly be required to determine the significance of the tabulated results.

As a general rule no composition of ~~classes into~~ an aggregate on the one hand, and on the other no 1) differentiation of an ~~aggregate into classes,~~ should be tabulated which does not ~~justify a percentage~~ distribution of the aggregate by classes.

A second principle to be observed in distributions is that the aggregate shall be, at least formally, 2) completely distributed. A partial distribution of the aggregate introduces a margin of error that may seriously impair the validity of the entire distribution. Where, for example, the replies to an inquiry are very incomplete, a percentage distribution based upon the total number of cases will understate the importance of each class, in proportion as the total embraces unknown cases. Where it may be fairly assumed that the unknown cases are distributed in the same proportion as the known to the several

classes, a percentage distribution may be based upon the known cases. The unknown cases should, in fact, generally be excluded from the base, where they are sufficiently numerous materially to affect the percentages.

3) A third principle involves the number of classes. As a general rule it is true that a distribution loses definition in proportion as the number of classes increases. This principle may be illustrated by the age distribution. A percentage distribution by single years of age involves approximately 100 classes and the number in each year normally decreases with advancing age. Reduced to percentages the proportion in each single year of age decreases from between two and three per cent in the younger ages to a small fraction of one per cent in the advanced ages. The differentiation is thus restricted to a very small range of variation, limited to approximately three per cent in the aggregate for the 100 classes. In other words, 100 classes must be differentiated by 100 percentages diminished by small fractions from three as a maximum. Moreover, the change from class to class over considerable life periods, those, namely, in which the rate of mortality is low, would be too slight to affect percentages worked to one place of decimals. For the fifty or fifty-five classes representing the ages above forty-five, the numbers in practically all of the classes would be represented by fractions of one per cent, and for all of these classes the variations from one year of age to a succeeding year would necessarily be defined in exceedingly minute frac-

tions. Such fine variations carry no differentiation of significance. The absolute decrease from year to year is a more comprehensible figure than is the minute modification of the percentage.

While percentage distributions by single years of age are sometimes introduced in statistical reports, a more effective method of indicating the relative numbers in the several ages is to relate the number in each year of age, not to the total of all ages, but to the number under one year of age. This method obviously gives a much wider range of variation for the relative numbers, the range being, in fact, a range of 100 per cent instead of two or three per cent, since the number under one year of age is always 100 per cent. On the basis of 1,000 or of 100,000 — the latter being the more common base — the number under one year is taken as 1,000 or 100,000, and the number in each other year is in the proportion to 1,000 or 100,000 that the absolute number in the given year is to the absolute number under one year of age. The resultant numbers it will be noted are not percentage, or per 1,000, or per 100,000 distributions of the total population, but are relations of the numbers in each single year of age to the number in one selected year of age.

To make an age distribution of the population effective on a percentage, or on any other basis, the number of classes must be reduced by combining single years into age periods of five or ten, or more years each. Such a distribution might for example show the percentage of the total population under fifteen years of age, fifteen to forty-four years,

forty-five to sixty-four years, and sixty-five years and over — or any other significant age grouping.

4 A fourth principle to be observed is involved in the nature of distributions, which relate all of the several classes to the aggregate of all classes. It follows obviously that distributions should be used chiefly, if not exclusively, in cases where this relationship of the part to the whole is more important than some other relationship, such as, for example, that of one class to another. In the age composition of a population, the relation of the number in each year of age to the aggregate population is much less significant than the relation of the number in each year to the number in other years, preferably to the number in the first year of age. Many other similar cases are encountered in statistical work.

6. Relations of Class to Class

Percentage distributions serve more or less effectively the double purpose of relating each class to the aggregate, and, through this relationship, the several classes to one another, but the relation of the classes to one another is not direct. The percentage foreign-born in the population, when related to the percentage native, indicates the relative numerical size of these classes, but it does not indicate relative size so clearly as a direct relation of the number in one class to the number in the other class, the number foreign-born, for example, per 1,000 natives. The percentages foreign-born, moreover, would not register a change in the relative number in the two classes so clearly as a direct re-

lationship, since both percentages are changing coincidently, one decreasing as the other increases, and since each class itself enters into the base to which it is related, and is to that extent related to itself in the percentages. An increase in the proportion foreign-born from twenty-five to fifty per cent, would increase the number foreign-born per 1,000 natives from 333 to 1,000, doubling the percentage being in this case equivalent to tripling the ratio of one to the other. A further increase of twenty-five in the percentage foreign-born would increase the ratio from 1,000 to 3,000, the ratio being tripled in this case by an increase of one-half in the percentage. The percentages of distribution indicate changes in the composition of the population, and the ratios indicate changes in one class relatively to the other. Similarly the percentage mulatto in the Negro population relates changes in the number of mulattoes to the total Negro population, a base which is itself partly mulatto and is, as regards its color composition, constantly changing in character, while the ratio of mulattoes to blacks relates changes in the number of mulattoes to changes in the number of blacks, a base which is, as regards its color composition, homogeneous, independent, and unchanging. The distribution of the population by sex, also, as has been noted, relates the number of males and of females to the total composed of males and females, while the sex ratio relates the number of males to the number of females, a relationship which is much more significant.

The employment of ratios showing the relation

of one class to another, involves a selection of one relation among all the different relations which might be shown, and an exercise of judgment which is not involved in the employment of percentage distributions. It is perhaps on this account that distributions are more commonly used than ratios in practical statistical work. The percentage distribution of the population classified according to marital condition into four classes, as single, married, widowed, or divorced, for example, does not involve any selection of relationships among the several classes, while the employment of a class ratio would involve a selection of one of four possible sets of relationships, as the one most significant: per 1,000 single, the number married, widowed, or divorced respectively; or per 1,000 married, the number single, widowed, or divorced; or per 1,000 widowed, the number single, married, or divorced; or per 1,000 divorced, the number single, married, or widowed.

The distinguishing characteristic of the class ratio is, therefore, that it involves a selection of one among several possible relations, and that it is specific in its significance; while the percentage distribution, since it relates each class to a heterogeneous total, is impartial as regards interclass relationships, which are all expressed indirectly through relationship to the total.

7. *Illustrations of Increases, Distributions, and Class Ratios*

The characteristics of increases, distributions, and class ratios may be illustrated by developing

such ratios under some simple assumptions such as are made in constructing the following table, namely, that a stationary population, assumed for convenience to be 100,000, composed originally entirely of foreign-born persons, is isolated and becomes, in consequence of births of native children and mortality in the original population, entirely native in the period of 100 years; the increase of natives by excess of births over deaths, and the decrease of foreign-born persons by mortality, in each decade being 10,000. Under these assumptions, certain percentage increases and decreases would develop in each decade, the composition of the population as regards nativity would be regularly modified, and the ratio of foreign-born persons to natives would diminish, as shown in the table. (See page 72.)

It will be noted that the native element, increasing by 10,000 in each decade, shows percentage increases ranging from 100 in the second decade to 11.1 in the tenth decade, the decline in the percentage being very rapid in the earlier decades, and relatively inconsiderable in the later decades. Although the native element increased in the first decade, no percentage increase can be determined for this decade since there is no base upon which to figure the percentage. The percentage decreases of the foreign-born element range from ten in the first decade to 100 in the tenth decade, the same percentages appearing, after the first decade, in the reverse order, as decreases for the foreign-born and increases for the native element. In each

HYPOTHETICAL POPULATION, STATIONARY, WITH INCREASING PROPORTION NATIVE

Year	Population			Decennial increase of native		Decennial decrease of foreign-born		Distribution		Number foreign-born per 1,000 native	Decrease of percentage foreign-born	Decrease in the number foreign-born per 1,000 native
	Total	Native	Foreign-born	Number	Per cent	Number	Per cent	Percentage native	Percentage foreign-born			
1800...	100,000	.....	100,000	.....	....	.....	....	....	100.0	....	....	....
1810...	100,000	10,000	90,000	10,000	....	10,000	10.0	10.0	90.0	9,000	10	....
1820...	100,000	20,000	80,000	10,000	100.0	10,000	11.1	20.0	80.0	4,000	10	5,000
1830...	100,000	30,000	70,000	10,000	50.0	10,000	12.7	30.0	70.0	2,333	10	1,667
1840...	100,000	40,000	60,000	10,000	33.3	10,000	14.3	40.0	60.0	1,500	10	833
1850...	100,000	50,000	50,000	10,000	25.0	10,000	16.7	50.0	50.0	1,000	10	500
1860...	100,000	60,000	40,000	10,000	20.0	10,000	20.0	60.0	40.0	666	10	334
1870...	100,000	70,000	30,000	10,000	16.7	10,000	25.0	70.0	30.0	429	10	237
1880...	100,000	80,000	20,000	10,000	14.3	10,000	33.3	80.0	20.0	250	10	179
1890...	100,000	90,000	10,000	10,000	12.7	10,000	50.0	90.0	10.0	111	10	139
1900...	100,000	100,000	.....	10,000	11.1	10,000	100.0	100.0	....	....	10	....

decade the percentage native increases and the percentage foreign-born decreases uniformly by ten.

The number foreign-born per 1,000 native decreases from 9,000 at the end of the first decade to 139 at the end of the ninth decade, there being no ratio of foreign-born to native at the beginning or end of the period, when the population is in one case 100 per cent foreign-born, and in the other 100 per cent native. An increase in the percentage native from ten to twenty reduces the number foreign-born per 1,000 native from 9,000 to 4,000 or by 5,000, and an equal increase of ten, from eighty to ninety in the percentage native reduces the number foreign-born per 1,000 native from 250 to 111, or by 139.

The proportional change in the ratio is the same in both of these cases, although the absolute decrease in the number foreign-born per thousand native, is much greater in the earlier decade, because the ratio itself is in that period much greater than it is in the later period. Incidentally it may be noted that the table illustrates the retardation of the rate of increase which, under a succession of fixed absolute increases of any amount, tends to approach zero as a limit.

The foregoing table is purely hypothetical, and is introduced simply to illustrate the character of the several derivatives under consideration. A concrete illustration of the development of such derivatives may be based upon the population returns of the Federal Census.

The population of the United States in 1910 and 1900 was distributed by color as follows:

Racial Class	Population	
	1910	1900
All classes	91,972,266	75,994,575
White	81,731,957	66,809,196
Negro	9,827,763	8,833,994
Other colored	412,546	351,385

For certain purposes this table as it stands might suffice, but it does not express any of the simpler relationships, of class to class, or of year to year, upon which in a great majority of cases the person consulting it would desire information. For his purpose the absolute number of Negroes in the United States in 1910 might be important only when related to the total population and reduced to a percentage of that total. He might want to know whether the proportion of Negroes in the population was larger in 1910 or 1900, or, stating the change in terms of increase, whether the Negro or White population had increased more rapidly during this decade. These simple percentages should obviously be included in the table.

The percentage of Negroes in the total population in 1910 is

$$\frac{9,827,763 \times 100}{91,972,266} = 10.7 \text{ per cent.}$$

The corresponding figures for 1900 are

$$\frac{8{,}833{,}994 \times 100}{75{,}994{,}575} = 11.6 \text{ per cent.}$$

These percentages indicate clearly that the proportion of Negroes in the total population was smaller in 1910 than in 1900; that per 1,000 population, the number of Negroes declined from 116 in 1900 to 107 in 1910.

The rate of increase in the Negro population from 1900 to 1910 is computed as follows: The actual increase in numbers was $9{,}827{,}763 - 8{,}833{,}994 = 993{,}769$. The percentage of increase is

$$\frac{993{,}769 \times 100}{8{,}833{,}994} = 11.2.$$

In the denominator is placed the figure for the year which serves as a base. In this case it was the Negro population in 1900.

When all of the percentages have been computed, the table may take the following form:

Racial Class	Population					
	1910	1900	Increase: 1900-1910		Percentage Distribution	
			Number	Per cent	1910	1900
All classes..	91,972,266	75,994,575	15,977,691	21.0	100.0	100.0
White........	81,731,957	66,809,196	14,922,761	22.3	88.9	87.9
Negro........	9,827,763	8,833,994	993,769	11.2	10.7	11.6
Other colored..	412,546	351,385	61,161	17.4	0.4	0.5

From this table it is apparent at a glance that the proportion of Negroes was larger in 1900 than in 1910, and the additional fact is brought out clearly that during this decade the White population had increased twice as rapidly as the Negro.

The population of the United States in 1910 and 1900 was distributed by sex as follows:

Year	Population	
	Male	Female
1910	47,332,277	44,639,989
1900	38,816,448	37,178,127

The absolute numbers are, however, in themselves incomprehensible, and barren of significance, except as they are related to one another by the introduction of derivative figures. A simple development of numbers is given below.

Year	Population					
	Both sexes	Male		Female		Males per 100 females
		Number	Per cent	Number	Per cent	
1910	91,972,266	47,332,277	51.5	44,639,989	48.5	106.0
1900	75,994,575	38,816,448	51.1	37,178,127	48.9	104.4

A few simple rules of statistical practice relating to percentages may be noted. In determining percentages it is generally advisable to calculate to two

decimal places, but in printing these figures one decimal place is sufficient and generally the margin of error in the data exceeds one-tenth of one per cent. Where the margin of error is large the decimal may be omitted altogether. Whenever the second decimal figure is less than five, the first decimal should be left unchanged, but whenever the second decimal is greater than five, the first decimal should be increased by one. Thus, 46.23 per cent would become 46.2 per cent and 56.38 per cent would become 56.4 per cent. It frequently happens that when a column of percentages computed in this way is added, the total will not be exactly 100 per cent. Thus, according to the Census of 1910, the White population of the state of New Hampshire was distributed as follows:

Native White—Native parentage	53.6%
Native White—Foreign or mixed parentage . . .	24.0
Foreign-born White	22.5

The total of these percentages is 100.1. Each one of these percentages is more nearly correct than any other figure carried to one decimal place would be. It is impossible to make these figures total 100 without forcing one of them. This is done at times in the following manner. Suppose we have the following percentages carried to two decimal places:

26.24
31.62
18.40
16.83
6.91

The total of these percentages is 100. The rule

having been adopted to print percentages only to one decimal place, it is apparent that the second decimal would be dropped in every case without increasing the value of the first decimal. If this were done, however, the total would become 99.9. The simplest way by which the total could be made 100 would be to increase by one the value of the first decimal which is followed by the largest second decimal. In this case the first number, 26.2 per cent, would become 26.3 per cent and the total would be 100 per cent. This has been accomplished, however, by forcing one percentage and making it less accurate than in the original form. It is generally better policy to leave each percentage as nearly correct as possible and not force the value of any of them, even if the total is not exactly 100 per cent. The number five may be considered as neutral. Where this is the second decimal in one or more cases in a column of percentages, it is justifiable either to leave the first decimal as it stands or to increase it by one in order to balance the total.

Percentages may be calculated either by division, logarithms, or the slide rule. Computation by means of the slide rule is more rapid than by either of the other methods, and gives results which are sufficiently accurate for ordinary purposes. Percentages should generally not be given where the base is less than 100.

8. Birth, Marriage, and Death-Rates

It is customary to compute birth, marriage, and death-rates upon a basis of 1,000 total population.

These are sometimes known as crude rates to distinguish them from refined or corrected rates. To compute the crude death-rate for the city of Washington for the year 1911 it is necessary to know the estimated population at the middle of the year, or on July 1, 1911, and the number of deaths during the calendar year 1911. The estimated population on this date was 337,475, and the number of deaths during the year was 6,304. The crude death-rate, therefore, becomes

$$\frac{6{,}304 \times 1{,}000}{337{,}475} = 18.7.$$

Although crude rates do not take into consideration differences in the sex and age constitution of the population, they possess considerable value. The composition of any population of considerable size changes but slowly, and changes in the crude death-rate reflect in a fairly satisfactory way changes in the healthfulness of the population. The difference between the crude birth-rate and the crude death-rate of a population represents the natural increase of the population.

Since the death-rates of children under five years of age, and of persons in advanced years, are higher than of those in the early age groups, we should expect a high crude death-rate in communities with a large proportion of infants or of those at advanced ages, whereas a low death-rate might be expected in communities with a large proportion of population in the middle-age groups. In the crude death-

rate no account is taken of differences in the sex and age composition of populations.

In order to obviate this difficulty, recourse is often had to specific death-rates. These rates are generally based upon a group of population limited by sex and age, since these characteristics more than any other factors influence the death-rate. Thus, specific death-rates might be computed for the population of two different countries distributed by sex and by five-year-age groups. It would then be possible to tell whether the specific death-rates were higher at all ages and for both sexes in one country or the other. The principal difficulty with specific death-rates is that they do not give in one figure a comparison of the mortality in the two communities. Such a figure is obtained by means of standardized or corrected death-rates.

In obtaining "corrected" death-rates the usual method is to select a standard population, definitely distributed into certain groups with respect to age, or age and sex; the specific death-rates of any area as computed for the same groups are then applied to corresponding subdivisions of the standard population, the result being the number of deaths which would have occurred in each group of the standard population had its death-rate been the same as that of the same group of the given area. The summation of the deaths that would have occurred in all the groups of the standard population gives the total number of deaths in the standard population corresponding to the observed specific death-rates in the given area, and the division of this total by the standard population yields the corrected death-rate.[1]

[1] Department of Commerce, Bureau of the Census, *Mortality Statistics, 1911, Twelfth Annual Report*, p. 20. For a more extended discussion of this subject consult this volume.

The International Statistical Institute in 1895 recommended that the population of Sweden in 1890 be accepted as the standard. There was no distribution according to sex and the population was divided into five age groups: Under one year, one to nineteen years, twenty to thirty-nine years, forty to fifty-nine years, and sixty years and over. The Registrar General of Great Britain has adopted as standard the population of England and Wales in 1901. By this method there are eleven age groups, each one distributed by sex. The corrected rates of any country are the ones which would have resulted from a combination of the specific rates for the various sex and age groups if the sex and age distribution of the populations had been identical with that of England and Wales in 1901. In this way most of the influence of differences in sex and age distribution are eliminated. For most purposes it may be noted the simpler method with but five age groups is satisfactory, but it is still true that the specific rate is the most significant. The corrected rates when thus computed usually differ but little from those obtained by the use of the English standard by which twenty-two specific rates are combined into one corrected rate. Any standard death-rate is obviously defective as a precise measure of mortality, since it does not show specific mortality by age. From a comparison of standard death-rates it would, for example, not be possible to determine to what extent a relatively high rate was due to a high rate of infant mortality. But, as has been noted, it facilitates comparisons of mortality in one

population with that in another, and is a much more precise measure of mortality than the crude death-rate.

It is evident that although the specific death-rates for each age group might be identical for two different populations, yet the differences in age distribution would cause a disparity in the crude death-rates. The following example will perhaps make this clear:

Age Groups	Negro			Foreign-born White		
	Percentage distribution	Specific death-rate per 1,000 population	Crude death-rate	Per cent distribution	Specific death-rate per 1,000 population	Crude death-rate
Under 5 years....	12.9	50.0	6.45	0.8	50.0	0.4
5-14 years......	24.4	4.0	.976	4.9	4.0	0.196
15-44 years......	48.1	6.0	2.886	59.9	6.0	3.594
45 years and over..	14.3	25.0	3.575	34.3	25.0	8.575
			13.887			12.765

It is assumed that for the Negro and foreign-born White population the specific death-rates are identical. Thus under five years the rate is fifty per 1,000; from five to fourteen years four per 1,000; from fifteen to forty-four years six per 1,000; and forty-five years and over, twenty-five per 1,000. The percentage distribution by age groups in 1910 is also given for these two classes of the population. Thus 12.9 per cent of the Negro population is under five years of age. The specific death-rate for those

of this age is fifty per 1,000. This will give 6.45 deaths of children under five years of age in 1,000 of total population. Continuing this process, we find the crude death-rate among the Negro population is about 13.9, while among the foreign-born White population it is 12.8. Apparently, then, the death-rate is considerably higher among Negroes than among the foreign-born White, and yet we assumed in the table that health conditions were absolutely identical for both groups and that there was no difference in the specific mortality rates. It is evident, therefore, that this apparent difference in crude death-rate is due entirely to differences in age composition. It is to eliminate this element of error that the standardized or corrected death-rate has been introduced into vital statistics.

Incidentally it may be noted that it is necessary to eliminate the influence of sex and age composition in the computation of many other rates, as well as in those relating to mortality. Thus, for instance, in a comparison of crime among native and foreign-born, it is a well-known fact that males of early middle life commit crime to a greater extent than females or males in the more advanced ages. The foreign-born population contains a larger proportion of adult males than the native-born population. Before we can compare the criminality of the native and foreign-born population, we must eliminate the differences due to sex and age classification. In other words, before we try to compare the relative frequency of different phenomena we must be certain that the composition of the base is identical in

all cases. One of the great causes of error in statistics is the assumption that other things are equal when, in most cases, other things are not equal.

The infantile death-rate is usually considered to be the number of deaths per annum of infants under one year of age per thousand children born during the year in question. This would give a correct rate on the assumption that births and deaths are recorded with equal accuracy. As a rule deaths are, in fact, more accurately reported than births. If, as years go on, an increasing proportion of births is reported and there is no change in the relative frequency of deaths, the decrease in the rate of infant mortality will be more apparent than real. The increase in the size of the denominator will be due not entirely to an increase in the number of births, but in part to an increase in the proportion of births reported. There is a lack of uniformity in the treatment of stillbirths in computing the infantile death-rate. In some cases these are included among both the births and the deaths, while in other cases they are returned separately and no notice is taken of them in computing the death-rates. The method followed should be explained in the text.

The crude birth-rate is the number of births per annum per one thousand of total population. This rate is open to the same objections that may be advanced against the crude death-rate. It does not take into consideration the composition of the population. Where a country has a large proportion of married women in the early and middle-age groups

we should naturally expect a high crude birth-rate. The refined birth-rate usually states the number of births per annum per 1,000 women between the ages of fifteen and forty-four. The refined legitimate birth-rate is the number of births per annum per 1,000 married women between the ages of fifteen and forty-four. The refined illegitimate birth-rate is the number of illegitimate births per annum per 1,000 unmarried women between the same ages.

The fecundity of marriage refers to the number of births per marriage. To obtain this figure from a census of population it is necessary to include only those marriages in which the probability of additional births in the future is quite small. In making a study of this kind the cases are usually distributed according to the age of the mother at marriage and the duration of the marriage. Further distinctions may be made with regard to race or nativity.[1]

The crude marriage-rate is the number of persons married per annum per thousand of total population. Another rate sometimes used is the number of marriages per annum per thousand of total population. This rate is, of course, just one-half the preceding rate. But here again no consideration is given to the age distribution of the population. The refined marriage-rate is the number of marriages per annum per thousand of marriageable persons. This in-

[1]Although material for study of this kind has been available in the Bureau of the Census for the censuses of 1900 and 1910, the only use that has been made of it has been in the report of the Immigration Commission on *The Fecundity of Immigrant Women.*

cludes not only all those who have never been married but all widowed and divorced over fifteen years of age. This varies from the denominator used in computing the refined birth-rate since increasing age places no limit to the possibility of marriage.

9. Density and Areality

In 1910, the 91,972,266 individuals comprising the population of the United States lived upon 2,973,890 square miles of land. By the simple method of division the population per square mile becomes 30.9. The advantages of the figure for density over the total figures for population and land area are evident. This figure is simple and can be remembered easily. Since 1790, when the first Federal Census was taken, the population has changed at each decade, and in addition the land area has changed at six different censuses. By substituting the population per square mile for two columns of cumbersome figures, variations in both dividend and divisor are obviated and it is possible to see at a glance the variations in the density of the population of this country.

Not only is it possible to trace the changes in the density of the population of one country at successive decades, but comparisons are easily made of the density of population of different countries at the same time. Another advantage lies in the ease with which comparative figures of this kind lend themselves to graphic representation.

Closely allied to density is areality. The land area of the United States in 1910 was 1,909,289,600

acres. Dividing this figure by the total population, the areality or acres per inhabitant is 20.7. As a rule, the density of population is used in comparing areas in which the population is congested, whereas areality is used to compare sparsely settled sections.

10. Heterogeneous Ratios

Birth-rates, marriage-rates, death-rates, rates of population, of density, and of areality are specific cases of a large class of ratios, each of which might be reduced to the number of one category per unit of another category.

Such ratios should be distinguished from averages. The density of population expressed as the number of inhabitants per square mile of territory, for example, is not an average of square-mile populations, although in this case the mathematical result obtained by dividing aggregate population by square miles of territory is identical with the result which would be obtained by adding together the specific square-mile populations and dividing by the number of square miles. The latter process would be the process of obtaining an average of square-mile populations; the former, is the process of determining the average population per square mile or ratios of population to area.

Similarly, the death-rate is the number of deaths per 1,000 population, not an average of deaths in specific population groups of 1,000, and in this case if population were divided into groups of 1,000 at the beginning of a year, and deaths in these several groups were recorded during the year and averaged,

the result would not be identical with the death-rate per thousand population. The population divided into groups at the beginning of the year would be reduced by mortality during the year, so that an average of deaths occurring in these groups would not be an average of deaths occurring in population groups constantly maintained at the full number of 1,000, which is the assumption of the death-rate, but an average of deaths occurring in groups constantly diminishing by mortality. The determination of averages involves various methods of weighting and of mathematical computation, and each average embraces a number of definite quantities which are averaged, the quantities averaged being necessarily of one category. The ratio relates two aggregates and expresses the numerical proportion of one to the other in terms of some common denominator.

In the case of every ratio the presumption is that it represents a significant relationship, and since the number of such relationships is indefinitely great, the number of ratios which might conceivably be constructed is, also, indefinitely great, and is in fact limited only by the availability of data. The crude death-rate, for example, is a relation of the number of deaths to population, but deaths may be classified by cause of death and deaths from each specific cause related to population. Deaths from a specific cause may be further classified by sex, age, race, marital condition, and occupation of decedent, and related to population classified also by sex, age, race, marital condition, and occupation. Further

classifications may be made by month of incidence of deaths from each specific cause, of decedents classified by sex, age, race, marital condition, and occupation; and, further, by geographical areas and by numerous other more or less significant characteristics of population and of deaths. No one of the classifications mentioned can be fairly characterized as insignificant, and yet in combination they yield a multiplicity of ratios which might easily exceed the number of deaths and even of population. The statistician is constantly confronted with this multiplicity of relationships, no one of which is entirely devoid of significance, and among which he must select those which are of prime importance.

In making this selection one principle must be carefully regarded, namely, that the significance of the ratio will be largely determined by the homogeneity of the terms of the ratio. In a population composed of two very diverse racial elements, for example, the crude death-rate, or relation of deaths to total population might have comparatively little significance, since a mortality rate for the two elements combined might not fairly represent either element in the population, being a purely artificial and insignificant figure. Such a condition obtains, in fact, in many southern communities in the United States in which the population is composed of Whites and Blacks. An unrefined mortality rate for such a population does not approximate the mortality for either element, and changes in such an artificial rate might be due to changes in the proportion of Blacks in the population, or to changes in mortality in

either element, or to any combination of changes in the two elements.

Where the population is sufficiently homogeneous, and is normally distributed by age and sex, the crude death-rate has significance; but for many purposes a refinement of the rate by age, sex, cause of death, or other characteristic, is essential. This refinement is essentially a process of eliminating heterogeneity in the terms of the mortality ratio, a process of securing as regards decedents and population a homogeneous group.

Ratios of this class are exceedingly diverse in character. While mortality tables relate deaths to population per 1,000, or, in case of specific rates by cause of death, on the basis of 100,000, frequently the ratio is a per capita figure—wealth per capita, public expenditure per capita, expenditure for teachers' salaries per capita of the children enrolled in schools, crop yield per capita, value of products per capita, acres of territory per capita. In other cases, population itself is related to a unit of some other category, as in the statement of population per square mile, number of inhabitants per owned home, population per mile of railway, population per family (*i.e.*, average number in a family obtained by relating aggregate population to aggregate number of families). Again, the ratio may relate independent classes to one another, as, for example, in the ratio stating number of children under five years of age per 1,000 women fifteen to forty-four years of age. Or the ratio may be entirely independent of any population figures, as in

the ratio of car mileage, train mileage, locomotive mileage, to ton mileage; or of crop yield to acres planted; or of acres in farms to number of farms. Any numerical relationship in fact which is of significance may be expressed in a statistical ratio.

From the foregoing comment the principles to be observed in constructing such ratios will be obvious. The relationship must have a specific significance and value, and the terms of the ratio must be sufficiently homogeneous to carry that degree of preciseness and definition which interpretation of the ratio requires. Unrefined ratios, such as the crude death-rate, may serve as indexes of change, but where the data warrant refinement, the resultant refined ratios give a value to the data which the crude ratio does not uncover. The refinement must, however, not exceed the capacity of the data, by implying significance which is not inherent in the data.

CHAPTER VII

AVERAGES

AS COMMONLY understood, the term average signifies that mean quantity found by adding together particular quantities of one category and dividing the sum by the number of particular quantities entering into it. This is the common significance of the term also in statistical operations, wherever an intermediate number or mean of a different character is not specifically indicated. In statistical terminology, however, the term average is employed as a general term to cover a variety of intermediates.

1. General Characteristics of Averages

Of these intermediates the principal ones are the simple or weighted arithmetic average, the geometric mean, the median, and the mode, each of which in any given case is a function of a group of particular quantities, and is in some degree typical of the individual quantities taken collectively. Each average, using the term in its broader sense, embraces the group, and provides a simple index or mathematical measure by which collective differences between groups of particular quantities, and collective changes from period to period in the particular

quantities of any given group may be summarily determined.

A summary comparison of the age composition of populations living in different communities, for example, or of the population of one community in different years, may be made by calculating the average age, or by finding the median age of the different populations. Such an average or median, although it does not indicate the age distribution of the several populations, embraces the individual ages of persons composing the populations, and tends to respond more or less definitely to changes in these individual ages.

While, however, any group of particular quantities may be averaged, and a mean or intermediate quantity determined for the group, no average so determined can be resolved back into the particular quantities which it represents. These separate quantities are not indicated by the mean any more specifically than is the shape and weight of a material object indicated by its center of gravity. The average age of a population cannot be resolved back into the individual ages averaged, since any given average age may represent an infinite number of age distributions. The arithmetic average age twenty years, for example, may represent a group of individuals each of whom is twenty years of age; or a group one-half of whom are ten and one-half thirty years of age; or a group one-third of whom are thirty and two-thirds fifteen years of age; or any other of the infinite number of age distributions that will give the designated average of twenty

In an investigation of the economic status of wage earners, for example, differences between the wage-earning power of individual laborers within defined groups may be of more significance than any averages or means that can be calculated for the groups. Where, on the other hand, for any homogeneous group of wage earners, minute and insignificant differences in individual wages develop, an average wage for the group is a convenient and significant simple measure of their wage earning capacity. In such cases the average facilitates comparison of one group with another, or of the condition obtaining at one period with that obtaining at another period, or of conditions obtaining at any given period in different communities.

In brief, the process of averaging, in one way or another eliminates particular differences in the quantities averaged, and this process obviously should generally not be employed in cases where the particular differences are important. The average is indicated as a convenient figure where the particular differences eliminated, or composed by the average, are relatively insignificant.

The specific character of each type of mean or intermediate figure is indicated in the several sections of this chapter in which the processes employed in determining the several types of averages are described in detail.

2. *Arithmetic Average*

(*a*).—*Simple Arithmetic Average.*—A simple arithmetic average, as has been noted, is the

intermediate or mean figure commonly indicated by the term average. It is formed by dividing the sum of the items by the number of items.

The practical uses of such an average are innumerable. A laborer employed under the piece-price system of payment may desire to know how much he is making a week in the long run. At certain seasons of the year work with him is plentiful, while at other seasons he is not employed full time, and as a result his pay envelope contains varying amounts from week to week. He may have been employed previously at a flat daily or weekly wage, and he finds that since he has been paid by the piece his pay is larger some weeks than it was under the previous system, but smaller in other weeks. The only way for him to judge whether the change in the system of payment has been beneficial to him is to strike an average for the time elapsed since he began piece work. To obtain this it is necessary for him to know the amount earned each week since the change went into effect. The average obtained may differ from his actual earnings in any week, but it will, nevertheless, measure his earnings per week in the long run more accurately than any item in the list, since it is determined by his aggregate earnings in a given period, and is the flat-rate per week equivalent to his varying weekly earnings. Under the piece wage system, he may have earned for twelve successive weeks the following amounts: $13.20, $14.10, $15.40, $15.30, $15.60, $8.25, $14.50, $15.35, $15.75, $15.80, $15.20, $14.90, amounting

for the twelve weeks to \$173.35. His average weekly earnings have, therefore, been

$$\frac{\$173.35}{12} = \$14.45$$

In computing a simple average each item enters singly into the aggregate of items, which are taken as being of equal importance, and are aggregated and resolved into an average without any manipulation by the process of weighting. The average is an arbitrary quantity, which may, but more commonly does not, correspond to any item in the list.

If a_1, $a_2 \ldots a_n$ represent the series of separate quantities to be averaged the formula for the simple arithmetic average is

$$\frac{a_1 + a_2 \ldots + a_n}{n}$$

(*b*).—*Weighted Arithmetic Average*.—The weighted differs from the simple average in that some or all of the items are taken in the aggregate more than once, each being thus weighted in the aggregate according to its relative importance.

A student in college, for example, may carry fifteen hours of work weekly, distributed as follows: Economics, five hours; history, five hours; Latin, two hours; French, two hours; rhetoric, one hour. Upon a scale of 100 his standing in these courses may be: In economics, ninety-two; in history, ninety; in Latin, seventy-eight; in French eighty-five; and in rhetoric, seventy-five. Weighting his standing in each subject by hours per week devoted

to the subject, the products given in the following table are obtained:

Courses	Hours per Week	Standing	Product
Economics......	5	92	460
History........	5	90	450
Latin..........	2	78	156
French.........	2	85	170
Rhetoric.......	1	75	75
	15	420	1,311

If an average standing for the student is computed without regard to the number of hours in the different courses, it is found to be

$$\frac{420}{5}=84,$$

the simple arithmetic average. But the student devoted more time to economics and history than to his other work, and these courses obviously should count for more than the others in determining his standing. To get a true statement of his general standing, his standing in the several courses should be weighted according to their importance, which, in this instance, is fairly indicated by the number of hours per week. If the sum of the products of the standing in each course by the number of hours per week devoted to the course, is divided by the sum of the hours per week in all courses, the result,

$$\frac{1311}{15}=87.4,$$

is a weighted arithmetic average. Changing the hours per week in the several courses will not affect the unweighted average, although such a change may affect materially the weighted average. If, for example, the student working the same number of hours per week devotes one hour a week to economics, two hours to history, five to Latin, two to French, and five to rhetoric, and attains the same standing in the several courses as is indicated above, the unweighted average standing will remain unchanged at eighty-four, while the average weighted by hours per week will drop to 80.5.

In computing a weighted average each item is multiplied by a number to give it weight proportional to its relative importance. The sum of these products divided by the sum of the weightings, in the foregoing illustration aggregate hours per week, gives the weighted arithmetic average.

This type of average is the one most commonly used in computing index numbers, where the weighting may be determined more or less arbitrarily. In a weighted average of prices, for example, an arbitrary system of weights may be applied to the prices averaged, so that the price of each commodity shall enter into the aggregate in proportion to its importance, in the aggregate either of the community's production, or of the community's consumption.

Weighting may be regarded as a method of correcting the bias or false weighting of a simple average. In the case of the student's general standing, for example, the simple and formally un-

weighted average was, in fact, heavily weighted, since, in the simple average, one hour devoted to rhetoric counted for as much as five hours devoted to economics, or to history, or two hours devoted to Latin or to French. This weight in favor of rhetoric is corrected by introducing as a factor the number of hours represented by each standing. In the formally weighted average, each hour of work is given a standing and a simple average taken for the full number of hours; the standing in each hour instead of in each course being taken as a separate quantity to be simply averaged with other hour-units of standing.

Similarly in an average for prices, the necessity for formal arbitrary weighting may be avoided by so making up the group of commodities, for which prices are quoted and averaged, that each class of commodities is duly represented according to its importance from the point of view of production or of consumption. Obviously, practically the same result is obtained by taking ten quotations of price of an important commodity with one quotation of an unimportant commodity, as is obtained by taking a single quotation for the important commodity ten times, and a single quotation for the unimportant commodity once.

The formula for the weighted arithmetic average is

$$\frac{(a_1 \times b_1) + (a_2 \times b_2) \ldots + (a_n \times b_n)}{b_1 + b_2 \ldots + b_n}$$

in which b_1, $b_2 \ldots b_n$ are the numbers employed

as weights for the quantities a_1, $a_2 \ldots a_n$ to be averaged.

3. The Geometric Mean

Another form of average, the use of which by statisticians has been limited almost entirely to the computation of index numbers, is the geometric mean. To compute this mean the items in the series are multiplied, instead of being added, as in the case of the arithmetic average, and the root of the product corresponding to the number of items is found. Thus, if the items are a_1, $a_2 \ldots a_n$, the geometric mean will be

$$\sqrt[n]{a_1 \times a_2 \ldots \times a_n}$$

Finding this mean generally involves, as a means of avoiding tedious computations, the use of logarithms. The natural number corresponding to the arithmetic mean of the logarithms of the items in the series is the geometric mean. It is true of the geometric mean as of the simple and weighted arithmetic average, that it is seldom precisely equal to any item in a series. The arithmetic average of a series is always somewhat greater than the geometric mean of the same series.

4. The Median

To compute the arithmetic average or the geometric mean, it is not necessary to arrange the items of the series in order of magnitude. For the computation of the median this arrangement is necessary. When the items have been thus arranged, if the number of items in the series is odd, the median

is the middle item, i. e., the item having just as many items preceding it as it has following it. If the series is composed of an even number of items, then the median falls between the two middle items. If these happen to be of equal magnitude, the median corresponds to them. If they are of different magnitude, the median is usually taken as the arithmetic average of the two middle items.

In most cases the median corresponds to one of the items in the series. If the series is regular, there will usually be a considerable grouping in the middle of the series, and even if the number of items in the series is even, it is quite probable that the number on each side of the median will be identical. It is not common, therefore, to have the median an arbitrary number. If the series is distributed symmetrically, the arithmetic average and the median will be found to correspond closely. If the frequency curve has a decided tendency to skewness, the median may possess but little value. For the average series the median possesses most of the qualities which give value to the arithmetic mean, and has the advantage of being much easier to determine, since little or no addition and division is required. It is possible to determine the value of the median by the graphic method.

The shape of the frequency curve may be approximately determined by finding the value of the quartiles or deciles of a series. The first quartile corresponds to the item midway between the top of the series and the median, which is the second quartile. The third quartile is the item half-way

between the median and the bottom of the series. The deciles are found by dividing the series into ten equal parts. Thus, if a series is composed of 137 items, the first decile is equal to the value of the thirteenth item plus seven-tenths the difference between the thirteenth and fourteenth items, and the other deciles are similarly located, the fifth decile corresponding to the median. The quartile and decile values indicate the distribution of the items about the median.

5. The Mode

The mode corresponds to the number which is most frequent in a statistical series and may therefore be called the point of greatest frequency. If the figures which make up the statistical series are charted in the form of a statistical curve, the mode will be the peak or highest point of the curve. It therefore differs from the other averages in that the mode always corresponds to some figure in the series. If a number be chosen at random, the choice will be more likely to fall upon the mode than any other number in the series. From the fact that the mode is the value of greatest frequency in the series, it has been given the name of the normal or typical value of the series.

The value of the mode is not at all affected by any values at the extreme ends of the series. Thus, the normal height of men is not affected by the presence of giants or dwarfs. After infancy has been passed, the normal age at death is the age at which more individuals die than at any other age.

The wage of the typical street railway conductor is the wage paid to the largest number of individuals engaged in this line of employment. A clothier manufacturing ready-made garments will make the largest number of suits for the normal or typical man, decreasing the number as the measurements depart from those of the normal man.

The principal advantages of the mode are its ease of determination, and the fact that it really corresponds to concrete members of the series. If the series is fairly symmetrical, the mode corresponds very closely to the arithmetic average and to the median.

6. Deviation from the Average

In certain cases the items composing a statistical series will vary but little from the arithmetic average of the series. In other cases the dispersion of the series or the variations of the individual items from the arithmetic average will be considerable, and it is essential in every case for the interpretation of the average to know whether the dispersion is small or great.

The average temperature of a certain city for a year means very little unless we are acquainted at the same time with the variations from this average at different seasons of the year. In the summer time most of the temperatures will be considerably above this average and in the winter time considerably below this average. In fact, if the curve of daily temperatures were constructed, it might be found that there were two points of greatest fre-

quency during the year, neither of which would approximate closely to the average. On the other hand, the height of adult males would show a much smaller dispersion than temperature, and one in which the average height would correspond closely to the height of greatest frequency. In other words, the curve for the height of individuals would be fairly uniform, with steep sides, whereas the curve for temperature would not be uniform and the sides of the curve would not be steep. The simplest way of showing dispersion is to state the arithmetic average together with the maximum and minimum values of the series. In the case of temperature, for example, we should have the average temperature for the year, and, what is more important, the extremes of temperature, which would indicate the range from the coldest to the hottest day.

It has been suggested that dispersion be measured by comparing half the difference between the upper and lower quartiles with the arithmetic average.[1] A more common method is to take the differences between each item of the series and the arithmetic average. All of the differences are considered as plus differences. Their sum divided by the number of items will give the average of the deviation of the items from the arithmetic average. The percentile deviation from the average is obtained by reducing the average deviation to a percentage of the arithmetic average.

The standard deviation is obtained by finding first

[1] A. L. Bowley. *Elements of Statistics,* p. 136, London, 1901.

the deviation of each item of the series from the arithmetic average. These deviations are then squared. The sum of the squares of the deviations is then divided by the number of items and the square root taken of this quotient. This square root is called the standard deviation.

The table on page 108 will show the method of computing the standard deviation from the average.

In the first column is given the number of males to 100 females in the different states in this country in 1910. In the second column is shown the deviation of the ratio in each state from the average for the country as a whole, which was 106. The third column headed "Frequency" gives the weighting or population of each state in hundreds of thousands. In the ordinary statistical series where each item has equal weight this column would be unnecessary, but in computing the sex composition of the entire country, a state with a population of 7,000,000 must be given seven times as much weight as one with a population of 1,000,000. The next column gives the squares of the deviation from the average. The last column is made up of the products of the two previous columns. The formula for the standard deviation where the items are weighted is

$$\sqrt{\frac{(fd^2)}{n}}$$

Since n is equal to the weighting of the items, the standard deviation therefore becomes

$$\sqrt{\frac{67,188.6}{919.8}}=8.5$$

POPULATION OF UNITED STATES IN 1910

State	Males to 100 females	Deviation from average d	Frequency f	d^2	fd^2
District of Columbia	91.3	14.7	3.3	216.09	713.1
Massachusetts	96.7	9.3	33.7	86.49	2,914.7
South Carolina	98.5	7.5	15.2	56.25	855.0
Maryland	98.9	7.1	13.0	50.41	655.3
North Carolina	99.2	6.8	22.1	46.24	1,021.9
Rhode Island	99.3	6.7	5.4	44.89	242.4
Georgia	100.1	5.9	26.1	34.81	908.5
New Hampshire	100.8	5.1	4.3	26.01	111.8
Virginia	100.9	5.1	20.6	26.01	535.8
Alabama	101.0	5.0	21.4	25.00	535.0
New York	101.2	4.8	91.1	23.04	2,098.9
Mississippi	101.6	4.4	18.0	19.36	348.5
Louisiana	101.7	4.3	16.6	18.49	306.9
Tennessee	102.1	3.9	21.8	15.21	331.6
Connecticut	102.3	3.7	11.1	13.69	152.0
New Jersey	102.9	3.1	25.4	9.61	244.1
Kentucky	103.0	3.0	22.9	9.00	206.1
Maine	103.2	2.8	7.4	7.84	58.0
Ohio	104.4	1.6	47.7	2.56	122.1
Delaware	104.6	1.4	2.0	1.96	3.9
Indiana	105.0	1.0	27.0	1.00	27.0
Missouri	105.1	0.9	32.9	0.81	26.6
Vermont	105.3	0.7	3.6	0.49	1.8
Pennsylvania	105.9	0.1	76.7	0.01	0.8
Arkansas	106.0	0.0	15.7	0.00	0.0
Iowa	106.6	0.6	22.2	0.36	8.0
Illinois	106.8	0.8	56.4	0.64	36.1
Michigan	107.3	1.3	28.1	1.69	47.5
Texas	107.4	1.4	39.0	1.96	76.4
Wisconsin	107.4	1.4	23.3	1.96	45.7
Florida	110.0	4.0	7.5	16.00	120.0
Kansas	110.0	4.0	16.9	16.00	270.4
Nebraska	111.2	5.2	11.9	27.04	321.8
Utah	111.5	5.5	3.7	30.25	111.9
West Virginia	111.6	5.6	12.2	31.36	382.6
Oklahoma	113.7	7.7	16.6	59.29	984.2
Minnesota	114.6	8.6	20.8	73.96	1,538.4
New Mexico	115.3	9.3	3.3	86.49	285.4
Colorado	116.9	10.9	8.0	118.81	950.5
South Dakota	118.9	12.9	5.8	166.41	965.2
North Dakota	122.4	16.4	5.8	268.96	1,560.0
California	125.4	19.4	23.8	376.36	8,957.4
Idaho	132.5	26.5	3.3	702.25	2,317.4
Oregon	133.2	27.2	6.7	739.84	4,956.9
Washington	136.3	30.3	11.4	918.09	10,466.2
Arizona	138.3	32.3	2.0	1,043.29	2,086.6
Montana	152.1	46.1	3.8	2,125.21	8,075.8
Wyoming	168.8	62.8	1.5	3,943.84	5,915.8
Nevada	179.2	73.2	0.8	5,358.24	4,286.6
Total			919.8		67,188.6

CHAPTER VIII

GRAPHIC REPRESENTATION

TO MOST persons except trained statisticians a long column or columns of figures are bewildering. Their meaning is not in any case apparent at a glance, and if the figures can be so illustrated as to tell their story more clearly and quickly, a happy effect is secured.

1. Utility of Graphic Representation

A person may be interested in the comparative area of Germany and the United States. With the square miles of territory in each country at his disposal, a reader could make his own computation, but how much more striking it would be to superimpose the map of Germany upon the map of the United States when both are drawn to the same scale, and then from his local knowledge of area and distances in this country obtain a much more intelligible idea of the area of Germany. The figures showing the value of the articles exported from the United States during the past ten years would show the effect of the European war, but fo most readers this would be much more apparent if the same information were put in the form of a diagram.

The value of this form of representation is particularly great where it is desired to show the

changes in two phenomena which are inter-related. Thus, if one curve were plotted to show the rise and fall in industrial prosperity in this country, and upon the same sheet were plotted the number of aliens entering our country annually, it would be found that both of these lines tended to rise and fall together. An increase in prosperity would soon be reflected in an increase in immigration to this country. The change in direction would probably come somewhat earlier in industrial conditions than in immigration. This would tend to show that the change in industrial conditions was the cause of the change in the rate of immigration. This would make the curve of immigration a good example of the curve of pursuit, and by moving backward the curve of immigration for a few months a higher coefficient of correlation would be obtained.

Both static and dynamic conditions may be represented by the graphic method, static charts being used to compare magnitudes at the same point or period of time, and dynamic charts to show changes in phenomena during a period of time. Since, however, most forms of graphic representation can be used to represent either static or dynamic conditions, this distinction possesses but little classificatory value in this connection. Forms of graphic representation may be classified simply according as points, bars, curves, areas, or maps are employed.

2. *Point Diagrams*

Intensity or frequency may be shown graphically by points or dots scattered over a given area. Thus,

if the average number of inhabitants per square mile in two communities were found to be in one case ten and in the other 100, the relative density of population in the two communities might be graphically represented by laying off two squares of equal size, and distributing upon one ten dots and upon the other one hundred. It will be obvious that a display of density in some respects more effective might be achieved by substituting for the squares a map showing the area of the two communities and distributing dots over the map areas, in proportion to the geographical distribution of population, each dot representing a given number of inhabitants. If such a system of dots were distributed over a map of the United States by county areas, the distribution and density of the population would be closely approximated. Similar maps might be prepared showing the distribution of immigrants or of racial elements, the yield of important agricultural crops (See *Thirteenth Census*, Vol. v, p. 734), or the distribution of farm animals. The United States Bureau of Immigration at one time published a map of Europe showing by the number of dots in the European countries the number of thousands of immigrants who had come to the United States from those countries during a year. In this way it was possible to get a very clear picture of those portions of Europe which were sending the most immigrants to us at that time.

It is to be noted, however, that such maps, although they give an impressionistic picture of geographical distribution, are generally difficult of

interpretation in numerical terms. It would be impossible to read off from a population map so prepared the specific densities for county areas, or even the approximate populations of the areas in the case of the more populous counties, in which the number of dots would be so great as practically to preclude the possibility of counting them. The preparation of such maps is, moreover, generally difficult, and some system of markings which indicates specific densities for the areas approximately is usually preferable to a proportionate distribution of dots (See section on maps on page 128).

In certain cases, however, a modification of the dot map is a most effective means of graphic representation, as, for example, where a society wishes to show for a city the location of cases of contagious diseases occurring during a given year, or to define areas of contagion as they develop during an epidemic. Pins of various colors are pushed into a map of the city to correspond with the location of cases of different diseases, and the grouping of these pin-heads in certain sections of the city shows at a glance the location of areas of contagion.

3. Bar Diagrams

A simple way of representing magnitudes graphically is by a series of horizontal or vertical lines of varying lengths proportional to the magnitudes to be compared. Any number may be represented graphically by a line drawn to scale, so that the length of the line is determined by the number represented; each unit of length in the line repre-

senting a unit or units enumerated in the number according to the scale arbitrarily determined upon. Obviously any unit of length may be taken as representing any number of units enumerated in the number, thereby determining the scale by which other lines may be drawn to represent coordinate numbers. The lengths of the lines stated in linear units will then be proportional to the numbers which they severally represent.

Lines so defined, when drawn to a common base line of zero value, constitute a simple form of statistical graph, and as the value lines are commonly drawn with appreciable breadth to distinguish them from the lighter structural and limiting scale lines, they are generally designated bars, and this type of diagram is generally designated a bar diagram. The width of the bar has, however, no numerical significance.

The lines or bars may be erected vertically at regular intervals upon a horizontal base line, or drawn horizontally from a vertical base line. Positive values may be represented on one side of the base line and minus values upon the other. The net gain or loss of population by interstate migration may, for example, be illustrated by a diagram in which bars are drawn to a common base, each bar proportional in length to the net gain or loss of a state, the net gains being represented on one side and the net losses on the other side of the base. Similarly, the excess or deficiency of males relatively to the number of females in the population classified by single years of age or by age periods,

may be represented by bars extended at right angles to a common base or axis in a positive or negative direction, each bar representing the excess or deficiency of males per 1,000 females in a specified age group. Commonly, however, the quantities represented are all positive, and the bars in such cases are all drawn upon one side of the common base.

In all bar diagrams it is essential that the scale be clearly indicated and that scale lines be drawn parallel to the base, so that the bars in the finished diagram are imposed upon a graduated surface, a sufficient number of scale lines being drawn to indicate approximately the numerical values of the several bars. These scale lines originate in an axis, or marginal line drawn vertical to the base, and graduated according to the scale of the diagram.

While in a properly constructed bar diagram approximate values for the several bars can be read by the scale, it is generally true that the scale cannot be drawn sufficiently fine to indicate to the eye the exact value of every bar. In some cases these exact values can be written into the diagram without impairing its graphic value. When this is not done the data should be presented in tabular form, since the bar diagram is more or less defective as a means of expressing values precisely. For precise accuracy, the numbers themselves are, in fact, the perfect means of expression. Statistical graphs should not be regarded as substitutes for statistical tables. A graph is not a substitute, but an illustration, and in all cases the data illustrated should be presented.

Bar diagrams are of small utility when the differences represented are, as developed by the scale, so minute as to be difficult of appreciation. Not infrequently, however, the appreciability and effectiveness of a diagram may be increased by making combinations in the data, or by changing the scale.

Where the bars represent the distribution of an aggregate, as of population by age, they may be drawn in contact and developed as areas, the aggregate area of the bars representing the aggregate number distributed, e.g., the aggregate population in an age diagram or pyramid, and the area of each bar being proportional to the number in the class represented by the bar. Where the bars taken collectively do not represent an aggregate, as, for example, where they represent populations at different dates, they should be drawn detached, and distributed at regular intervals along the base. The intervals may, however, be so reduced in width as to bring the bars into very close proximity to one another.

Where a diagram is constructed to show the distribution of a single aggregate into classes, the absolute numbers in the several classes may be used and the scale adapted to comprehend these numbers, but in cases where the distribution of one aggregate is to be compared with the distribution of another, it is frequently necessary to reduce both distributions to a percentage, or per thousand basis, in order to make the comparison effective. A diagram showing, for example, for any year the distribution of deaths in the White and Negro population classified

by age would, if based upon the absolute number of deaths occurring at each age in each class, be ineffective because the number of White deaths greatly exceeds the number of Negro deaths, and a scale adapted to the number of White deaths would not develop satisfactorily the distribution of Negro deaths. In this case a diagram constructed to show not the absolute number of deaths at each age, but the deaths at each age per 1,000 deaths at all ages, would develop the distribution of deaths by age equally for the two classes.

Different relationships may be developed by various combinations, groupings and arrangements of bars. In a diagram representing specific mortality by age, differences between racial classes may be made apparent by grouping the bars for the several classes under years of age or age periods. Similarly, changes in specific mortality by age in any one class from year to year may be shown by grouping the bars for the several years under the ages, and differences from one community to another by grouping the bars for different communities under the ages. Such groupings have each of them a special significance and value.

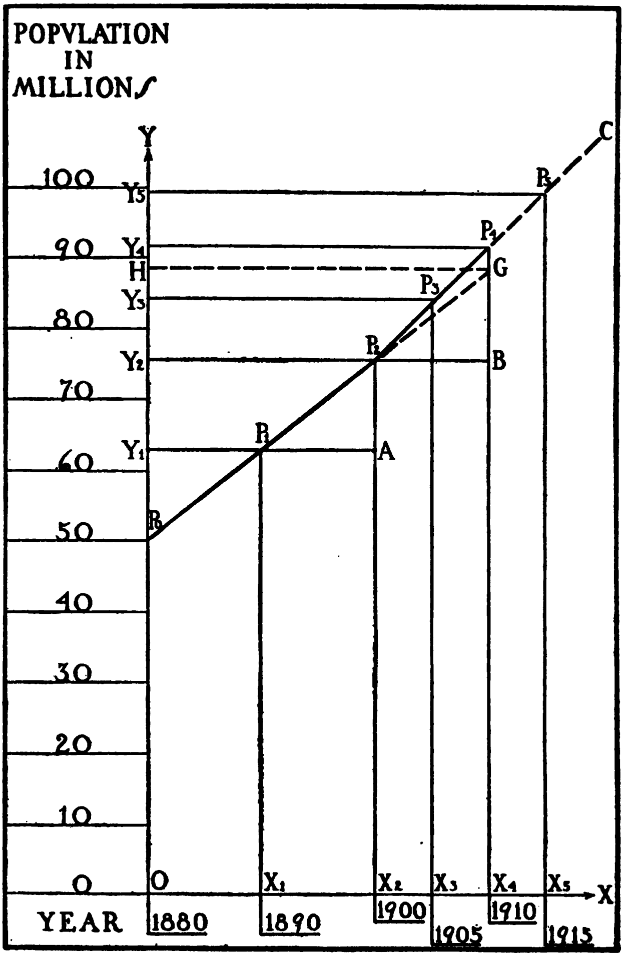
Where two systems of bars are developed from a common base they may be drawn on opposite sides of a central axis. A common instance of such a diagram is found in the age pyramid, which shows the distribution by age of males upon one side of the axis and of females upon the other. In a similar manner the age distribution of Whites and Negroes may be shown, or of native and of foreign-born

persons, or of the urban and the rural population, or of any other two classes. In such arrangement large differences between classes become more or less apparent in the lack of symmetry in the resultant graph with reference to the central axis.

4. Curves

The growth of population by decades may be illustrated by a simple bar diagram which is schematically a series of lines proportional in length to the population as enumerated in the decennial years. To draw such a system of lines distributed at equal intervals upon a common base, it is necessary to locate a series of terminating points above the base by reference to a graduated scale of population. If, instead of erecting perpendiculars from the base to these points, straight lines be drawn connecting the points with one another, a crude statistical curve is produced. Each of the points joined, if it has been properly located by reference to graduated scales of population and of decades or years, represents a specific period of time elapsed and a specific population related to that period. Each point is located by its coordinates of population and time elapsed.

The construction of such a curve is illustrated in the diagram on page 118, in which the curve P_0P_4 represents the growth of population in the United States during the three decades 1880-1910. On the perpendicular Oy equal spaces represent equal increments of population, and on the base line Ox equal spaces represent equal increments of time,


POPVLATION IN MILLIONS
Y
C
100
Y5
P5
P4
90
Y4
H
G
P3
Y3
80
P2
Y2
B
70
P1
Y1
A
60
P0
50
40
30
20
10
0
O
X1
X2
X3
X4
X5
X
YEAR
1880
1890
1900
1905
1910
1915

measured in each case from the point O. The space O_{x_1} is taken to represent the ten years 1880-1890, and the sequent spaces, $x_1 x_2$ and $x_2 x_4$, the decades 1890-1900 and 1900-1910. The curve originates in the point P_0 in the line Oy, which is the point in the line Oy corresponding to the population in 1880 (50,155,783). The points y_1, y_2, and y_4 correspond similarly to the population as enumerated in 1890 (62,947,714), 1900 (75,994,575), and 1910 (91,972,266). The points x_1, x_2, and x_4 correspond to the years 1890, 1900, and 1910, and the distance of these points from O represents periods elapsed subsequent to 1880 of ten, twenty and thirty years respectively. The points P_1, P_2, and P_4 are located by drawing lines from the y points parallel to the base Ox and erecting perpendiculars from the x points parallel to Oy. The point P_1, for example, is the point of intersection of the perpendicular erected at x_1 and the line drawn through y_1 parallel to Ox.

So located by their respective coordinates, the points P_0, P_1, P_2, and P_4 represent graphically the enumerated populations for the years 1880, 1890, 1900, and 1910, and it will be obvious that any point in the line or curve connecting these points, P_3 for example, represents presumptively a specific population and a specific period or date. On the presumption that population is increasing by equal increments from year to year, the population in the middle of the decade 1900-1910 may be determined by erecting a perpendicular at the point x_3, midway between x_2 and x_4, and drawing a line from the

point of intersection with the curves, parallel to Ox, to the line Oy. On the same assumption, the population in any year subsequent to the enumeration in 1910 may be determined by extending the line P_2 P_4 in the direction of c, and erecting a perpendicular from x_5, the point corresponding to the date for which the population is to be determined.

If, however, population is increasing by unequal increments from year to year, this method will not give a correct result. Such an extension of the line $P_1 P_2$ as shown on the diagram would have indicated erroneous populations for the years intervening between 1900 and 1910, because population was increasing in this latter decade by annual increments greater than those of the preceding decade. The broken curve made up of straight lines connecting the points determined from the date of enumeration, proceeds on the assumption that the rate of increase per cent for population during each decade is decreasing regularly from year to year, the decrease in the rate being just sufficient to prevent an increase in the increments of population per unit of time. If this is not the case, that is to say, if the rate of increase per cent is constant, or is increasing or decreasing regularly so as to produce varying increments of population per unit of time, the correct population curve will not be in any section of it a straight line. The straight line connecting P_1 and P_2 distributes the aggregate increase for the decade 1890-1900, represented by the line aP_2, to the several years, months, or days of the

decade, so that the increase in any time interval taken in the decade exactly equals the increase in any other time interval of equal duration, the increase in the first month of the decade being precisely equal to the increase in the last month. In the diagram the greater steepness or inclination to the base Ox, of the line $P_2 P_4$ as compared with the line $P_1 P_2$, indicates that the curve $P_1 P_2 P_4$ is only approximately the true population curve for the period 1880-1910. If the number of points established by enumeration in this period were increased, a closer approximation to the true curve would be given by straight lines connecting these points. The process of drawing a true curve through the points established by enumeration requires separate treatment.

The most common use of curves in graphic representation is to represent by a line or curve constructed as described above, the changes in some phenomenon during a period of time. The drawing of such a line is simplified by the use of cross-section plotting paper laid off in squares or rectangles by light blue lines, with, as a rule, every fifth or tenth line drawn heavier than the others to facilitate the placing of points. When a photographic reproduction of the chart drawn upon such paper is to be made, those lines should be drawn in black which are to be reproduced in the photograph, since the pale-blue cross-section lines will usually completely disappear in the photograph. Years or months should generally be arranged along the base of the chart, with the earliest date at the left. The

scale of magnitudes should be in a column at the left, arranged in ascending order of magnitude, beginning with O at the base line. When the magnitude for any particular year is to be placed, the perpendicular line corresponding to the year is followed until it crosses the horizontal line corresponding to the magnitude, at which point a cross or dot is made upon the paper. When the magnitudes for each year have been placed in this way upon the paper, the broken straight line connecting the points is drawn in considerably heavier than the lines forming the squares or rectangles.

Several different facts may be shown upon a single chart by using different kinds of lines, composed of dashes, long, short or alternating long and short, of small crosses or circles, or any combination of distinguishing marks. It is generally unwise, however, to place more than eight or ten lines upon the same diagram, and even this number may be confusing to the eye which may find difficulty in following individual lines, especially in cases where the lines are bunched together in some sections, or closely interfere throughout, or frequently cross one another. In hand work, it is possible to use inks or crayons of different colors in the lines, but for ordinary printing it is better to use lines of different formation.

For certain purposes it is preferable to use paper with logarithmic ruling in either one or both directions, and Professor Fisher[1] has recommended in

[1] *The Annalist*, March 9, 1917.

certain cases the use of a percentage plotting paper, where —

> . . . . a given elevation of any one point above another represents a given percentage excess of the first number above the second; so that when two curves are parallel they are really following the same percentage rise or fall, and when one curve is the steeper it is really rising at a higher percentage rate.

True curves may be substituted for the broken straight lines, on the assumption that in these broken lines, which constitute curves of error, the irregular shapes, breaks, or angles are caused by the fact that the number of observations is insufficient to produce the symmetry or regularity characteristic of the true curve. When the points of a curve have been located, so far as indicated by the data, a true curve may be drawn mechanically by the use of a rule, so constructed as to provide in some section of its variously curving edge a perfect mechanical fit for any three consecutive points located.

Broken lines representing magnitudes may be drawn about a center. In this style of graph, a circle is divided by radii into as many sectors of equal size as there are items to be represented. The magnitudes are represented by measuring from the center along the radii distances corresponding to the magnitudes under observation, and placing crosses or dots in the proper locations on the radii. One coordinate of each point is thus measured on the circumference and the other on the radius of the circle. A broken line is then drawn connecting the points on the radii. If, for example, twelve radii

be drawn dividing a circle into twelve equal parts, one for each of the twelve months in the year, the average temperature for the several months may be marked off on the radii, measuring from the center, and when the points indicating average temperature are connected, the broad changes in temperature during the year become apparent. A similar form of graph is in use for recording automatically changes in temperature from day to day. A circle or dial divided into thirty sections slowly revolves in contact with a pen, the length of whose arm varies with the temperature, making a continuous record upon the revolving dial of changes in temperature throughout every day for a month. The varying length of the arm measures one coordinate, and the revolution of the dial the other coordinate, of each point in the temperature curve, drawn by the pen.

5. Surfaces

In the surface diagrams numbers are represented graphically not by proportional lengths of lines, but by proportional areas, which are commonly developed as rectangles, triangles, or circles, although other forms are used infrequently.

Any series of numbers may be represented by a series of similar geometrical figures, as is the case where squares, equilateral triangles, or circles are used, or by a series of dissimilar figures, as is the case where rectangles, of uniform width and varying length, or of uniform length and varying width, are used, or triangles of uniform base and vary-

ing altitude. Generally comparisons of areas are difficult where geometrically similar figures are employed, since in such figures all homologous dimensions of the figures are variables, varying in proportion to the square roots of the magnitudes represented by the areas.

In the display of area diagrams, space may frequently be economized by superposition of areas, small squares being drawn within larger, triangles drawn to a common base, and circles to a common center.

The construction of a series of similar figures of any shape to represent by proportional areas a series of magnitudes, involves application of the general principle that areas of similar surfaces vary in proportion to the squares of homologous dimensions.

Rectangles of equal base, placed side by side, with altitudes corresponding to the magnitudes to be compared, or rectangles of equal height and varying length, arranged one above another, constitute the simplest form of representation by surfaces. In representation by rectangles it is generally advisable to retain one dimension unvarying, since, as has been noted of similar figures generally, where both the base and the altitude vary, accurate estimation of areas is difficult.

The arrangement of these rectangles is not uniform in statistical practice. Where the horizontal length is variable, it is sometimes the practice to have the earliest year in a statistical series at the top, following down with the later years. In United

States census practice the latest year is usually placed at the top, and the earlier years underneath this. When these rectangles are placed on end, one most commonly finds the earliest year at the left and the latest year at the right of the series.

Upon rectangles used in this way it is possible to distinguish the distribution of different classes of items entering into totals. Thus, the population of the New England states might be represented by the rectangles of varying length, upon each of which the marital condition of the population might be distinguished by subdividing the areas, and using different forms of hachure shading to designate the population returned as single, married, widowed, and divorced, respectively. Solid black, white in outline, diagonal hatchings in different directions, and cross-hatchings, heavy and light, stripes and checks, are some of the common markings available for distinguishing surfaces.

In some cases rectangles are placed one upon another as in the construction of an age pyramid. Those under ten years of age might, for example, be represented by the lowest rectangle, upon which is placed a second rectangle representing those from ten to nineteen years of age. In a stationary population where uniform age periods are represented, each rectangle would be somewhat shorter than the one directly beneath it. An added distinction according to sex is shown if the males are placed to the left and the females to the right of a perpendicular line dividing the chart.

There are comparatively few cases where the triangle or the circle is so effective a means of graphic representation as the rectangle. Generally, comparison of triangular and of circular areas is difficult. The distribution of the population of a country according to wealth is sometimes represented by a triangle, the relatively numerous class of small property holders being represented by the large area at the base of the triangle, and the few of great wealth by the small area at the top. A diagram of this nature is, however, apt to be misleading.

Circles are sometimes used to represent the frequencies or magnitudes by their areas. Since the area of circles varies as the square of their radii, it is common in drawing circles to represent a series of numbers to take the square roots of the numbers to be compared, as the radii. The eye is, however, unable to measure accurately magnitudes represented in this way.

A more effective use of the circle is that in which radii are drawn dividing the circle into segments, corresponding to the magnitudes to be represented. Thus, a circle may be divided into four parts, distinguishing the native White of native parentage, native White of foreign parentage, foreign-born White and Colored. If two circles of equal size were drawn to compare the distribution of the urban and rural population by color, nativity, and parentage, one of the circles would represent the urban population and the other the rural, and by properly charting the segments these distinctions

between the two groups would be brought out clearly.

6. Stereograms

It is possible to represent three variables by the use of stereograms or solids, but these are difficult to construct and to reproduce in a volume. In its simpler form the stereogram may be briefly described as a projection of two systems of bar diagrams, representing in combination cross distributions of frequencies.

7. Maps

The ordinary use of the map for graphic representation is that in which the major or minor political divisions of a country have been outlined and the areas shaded according to the geographical intensity of the phenomenon represented. Thus, the percentage of illiterates in the population of the United States ten years of age and over might be shown for the different states by leaving in white those states in which illiteracy amounted to less than one per cent, hatching in light those in which it was from one to three per cent, hatching in darker those in which it was from three to five per cent, and so on, the area of states with an illiteracy of twenty-five per cent or more being filled in entirely black. This use of the map lends itself freely to the comparison of many kinds of phenomena, and is one of the clearest ways for showing sectional differences. The shading of the map should graduate from light to dark in proportion as the intensity

of the phenomenon increases. If colors are used, differences of shade in any single color should represent differences in intensity; and differences in color should generally represent intensities of different orders. The number of shadings should not be so great that the eye cannot easily distinguish them from one another.

8. Improper Use of Diagrams

In the use of diagrams there are many dangers to be avoided. The most common of these have been noted in a preliminary report published by the Joint Committee on Standards for Graphic Presentation. They are as follows:

The general arrangement of a diagram should proceed from left to right.

Where possible, represent quantities by linear magnitudes, as areas or volumes are more likely to be misinterpreted.

For a curve the vertical scale, whenever practicable, should be so selected that the zero line will appear on the diagram.

If the zero line of the vertical scale will not normally appear on the curve diagram, the zero line should be shown by the use of a horizontal break in the diagram.

The zero lines of the scales for a curve should be sharply distinguished from the other coordinate lines.

For curves having a scale representing percentages, it is usually desirable to emphasize in some distinctive way the 100 per cent line or other line used as a basis of comparison.

When the scale of a diagram refers to dates, and the period represented is not a complete unit, it is better not to emphasize the first and last ordinates, since such a diagram does not represent the beginning or end of time.

When curves are drawn on logarithmic coordinates, the limiting lines of the diagram should each be at some power of ten on the logarithmic scales.

It is advisable not to show any more coordinate lines than necessary to guide the eye in reading the diagram.

The curve lines of a diagram should be sharply distinguished from the ruling.

In curves representing a series of observations, it is advisable, whenever possible, to indicate clearly on the diagram all the points representing the separate observations.

The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top.

Figures for the scales of a diagram should be placed at the left and at the bottom or along the respective axes.

It is often desirable to include in the diagram the numerical data or formulae represented.

If numerical data are not included in the diagram, it is desirable to give the data in tabular form accompanying the diagram.

All lettering and all figures on a diagram should be placed so as to be easily read from the base as the bottom, or from the right-hand edge of the diagram as the bottom.

The title of a diagram should be made as clear and complete as possible. Sub-titles or descriptions should be added if necessary to insure clearness.

CHAPTER IX

CORRELATION

AT ANY given time, present conditions obtaining in society have been determined by the whole complex of natural and social forces which have been operative in the past, and it is by taking account of these conditions at regular intervals that the mass effect or resultant social tendency of these forces, that is to say, the direction and amount of social progress in successive periods, is determined.

The age classification of the population, for example, as set forth in the Population Census, is determined directly by factors of natality and of mortality operative during a period of approximately one hundred years prior to the taking of the census, and indirectly by every influence, however remote in time, which has in its ultimate consequence in any way affected the natural vitality of the population. The present age distribution of the population, in fact, sums up the natural progress of the race from its most remote origins. Similarly, the accumulated wealth of a community at any given time embraces and sums up the technical development of wealth-producing processes, and every agency of social economy in the past.

No social tendency is the simple and unrelated effect of a single cause. On the contrary, each

social tendency is more or less intimately associated as a composite effect or resultant, and as a reacting and original cause with every other social tendency. This is as true in the social world as it is in the physical, in which the direction and velocity of movement of any falling body is the resultant of the gravity pulls upon it of the whole universe of matter, the resultant being constantly modified by the falling of the body itself. These gravity pulls may equally well be regarded as influences originating in the falling body itself, which literally attracts the universe, and tends to cause a displacement of other matter precisely equal to its own displacement, in mass, time, and space. The falling body typifies social phenomena. Every social tendency is a sort of gravitation, a response to social gravity pulls, and at the same time an influence causing general social displacement and readjustment, without any definite limit.

In a word, social phenomena are universally correlated, interrelated and mutually reactionary, and although this universal correlation cannot be established or proved statistically, it is, nevertheless, a working hypothesis upon which statistical inquiry proceeds. It follows that every social phenomenon is, on the one hand, an effect, and on the other a cause, of other phenomena of equal social value. A complete analysis separating out all the influences acting and reacting in any social tendency is impossible, and frequently it is exceedingly difficult to isolate even the immediate principal influences. To take a single illustration entirely typical of social phe-

nomena in general, the decline in the birth-rate would appear to be one simple effect of a complex of social forces, vaguely comprehended in the term, civilization. To what extent, if at all, this decline is due to economic, or physical, or moral causes cannot be determined from any available data. The decline may be correlated with the accumulation of wealth, with changes in the price of food, with the increase in density of population, with the decline in the death-rate, with changes in the earning power of labor, with the entrance of women into wage-earning employments, with the development of certain social vices, with the consumption of articles of luxury, or with any other index of social progress or of social degeneration. In each case, if data are available, some degree of correlation will be found, either negative or positive, since the decline in the birth-rate is undoubtedly somewhat affected by each of the influences specified, and by many other influences besides. But since all of these influences are acting, coincidently, the precise effect of any one influence cannot be isolated by any manipulation of available data. It is equally impossible to determine the social reactions upon the decline of the birth-rate itself,—to determine, for instance, the ultimate social consequences of the decline in the size of families, which is involved with a decline in the marriage-rate and its social consequences; or of the decline numerically of certain social classes, and conceivably of population itself. It has been contended that decline in the birth-rate was a cause contributing largely to the

decline of the Greek and Roman and of other civilizations in the past, and that in societies today the decline affects principally the upper classes, so that population blights out at the top. On the other hand, restriction of the number of children born is regarded as a means of insuring a higher standard of living in the community and a more wholesome environment for the children born.

The increase of population in any community may be accurately determined by enumeration at regular intervals, but this increase will in every case represent a complex of social causes which embrace all of those influences affecting the mortality, the natality, and the migration of population. As wealth accumulates, it may be noted that the death-rate and the birth-rate decline, but coincidently with the accumulation of wealth, there will have taken place other social changes, which may have affected the vitality of the population. In recent decades both the birth-rate and the death-rate have declined in practically all civilized communities, and an effort has been made to determine the principal causes of this decline.

While it is generally true of social phenomena that they cannot be completely analyzed, unless in any given case some relationship or correlation is established, one primary purpose of statistical inquiry has not been achieved. So long as the social tendency remains uncorrelated, it is an isolated phenomenon which cannot be intelligently comprehended. It constitutes an unsolved statistical problem.

When confronted with such a problem, the statistician resorts to methods of correlation for the purpose of establishing a presumption that an immediate relationship of cause and effect does or does not obtain between two observed and statistically determined social tendencies.

Although such a relationship cannot be proved or disproved by any statistical method, where a high degree of correlation is observed and determined, the probability that a relationship of cause and effect obtains is frequently so great as to amount to a demonstration.

It is to be noted, however, that even a high degree of correlation may represent mere coincidence of phenomena. In a community in which the birth-rate is declining many other tendencies will be in evidence which cannot be presumed to have any appreciable effect upon the birth-rate. A considerable degree of correlation might be found between the decline in the birth-rate and the extension of railroad mileage, in a community over a long period of time, or between the birth-rate and the increase in the number of letters sent through the post-office. These correlations would be accidental and would not obtain in the annual fluctuations of births, mileage, and letters. It is necessary to be constantly on guard against such accidental correlations in statistical work, since the statistical process, being purely mathematical, does not discriminate between a causal and an accidental relationship.

As determined statistically, correlation is merely coincidence of social phenomena. Where two social

tendencies are observed to move coincidently, or where a presumption of relationship obtains a priori; and the two tendencies are susceptible of quantitive measurement, the closeness and degree of coincidence can be mathematically determined and a coefficient of correlation calculated in accordance with formulae which have been generally accepted by statisticians as giving an accurate measure of the closeness of coincidence.

More generally, however, correlations are simply stated without the use of formulae, as coincident increases or decreases, or they are displayed graphically in simple curve diagrams. The number of immigrants arriving each year in a given period may be directly compared with the number of commercial or industrial failures, with the number unemployed, with the volume of output of certain basic industries, with the volume of bank clearings, with some index of price changes, or with any other statistical measure of industrial conditions, with a view to noting coincident or divergent changes, the presumption being that the volume of immigration is affected more or less by fluctuations in industrial activity.

An illustration of a display of correlation by a very simple diagram is given in the *Statistical Register of South Australia for 1915-16.*[1] This diagram, covering a period of twenty-six years, comprises four curves showing, respectively, acreage under wheat, production of wheat, average yield per acre, and mean rainfall from April to

[1] Part III, Section I, p. 11.

November, by years. It is noted in the text accompanying the diagram that "A remarkable similarity is noticeable in the curves for rainfall, total production, and average yield per acre." It would be possible by the use of conventional formulae to calculate the coefficient of correlation between the rainfall and the yield per acre, and between the rainfall and the total production. A comparison of these coefficients would show whether the correlation of rainfall with yield per acre was closer and more immediate than the correlation with total production. The general coincidence of the increases and decreases in rainfall and in yield per acre and total production is, however, apparent in the diagram, which, in fact, gives information which could not be expressed by any simple coefficient of correlation, the specific changes from year to year being more significant than any summarization of the degree of coincidence in these changes.

In many lines of statistical inquiry, however, the coincidence of change or fluctuation is of prime importance. This is particularly the case where reasonable doubt attaches to the presumption of any important or direct interrelationship, and where the purpose of the inquiry is to establish the fact and the degree of correlation. In such cases an accurate measure of correlation is essential. Economists have been in very general disagreement in their several estimates of the factors affecting, for example, the general level of prices in a community, including the effect upon prices of fluctuations in the quantity of money, checks, or other credit

instruments in circulation, in the quantity of gold produced, in the rate of interest or discount, or foreign exchange, in the volume of bank reserves and deposits, and in such intangible factors as business distrust or confidence, monopolistic combinations, and manipulation of the market. In such cases where the degree of correlation between any two observed phenomena may not be apparent in a simple tabulation and charting of the data, some refined measure of correlation may be required.

Such a measure is, perhaps, of more obvious scientific value and of greater practical utility in those lines of scientific inquiry in which the data are characteristically more precise and simple than are the data of the social sciences. In social phenomena the correspondence of tendencies must generally be large and obvious in order to be entirely convincing, and where a refinement of measurement is required to uncover the correspondence, the refinement will generally impose upon the data an unwarranted degree of accuracy and of directness and certainty of inter-relationships in social tendencies. It may be shown, for example, that the tendency to commit crime varies with age, rising to a maximum in a certain age period and declining in more advanced ages, while the tendency to suicide increases with advancing age. These tendencies might be apparent in the statistics of crime and of suicide covering a period of years, under varying economic conditions, and in different communities and population classes showing different rates of criminality and of suicide. For the entire mass of

data classifying crime and suicide by age, coefficients of correlation might be calculated which would measure accurately the correlation between age and crime, and between age and suicide, and determine formally that the correlation between age and crime was greater, or less, than the correlation between age and suicide. But this whole process of refinement would be imposed upon data showing rates of crime and suicide affected by other factors than age. Coincidently with growing old, many conditions of life change, and some of these external changes may be important factors in developing or suppressing the tendency to crime or to suicide. The mathematical process measures the apparent correlation precisely, but it does not, of course, remove the uncertainty attaching to the original data as regards other conditions of the rates than merely growing old. Until these other conditions can be subjected to the same degree of refinement as the age conditions, our knowledge of the effect of the age factor, even after mathematical calculation of the correlation measures, will still be fairly expressed in general terms, rather than in precise numerical terms.

The graphic method of representing correlations does not require any special treatment. Two curves, representing supposedly related phenomena, as, for example, mean rainfall and average yield of wheat per acre, over a series of years, are brought into proximity by an adjustment of scales, and the extent to which fluctuations in one curve correspond with similar positive or negative fluctuations in the other,

is noted. The correspondence, or lack of correspondence, can be observed and described in general terms, but such a graph, although it displays the correlation, does not measure it.

If it is proposed to measure the correlation precisely, the data must be subjected to rather involved mathematical treatment. The mathematical derivation of the formula for determining the coefficient of correlation is set forth in detail in the more comprehensive systematic treatises dealing with the mathematical aspects of statistical computations.[1] This derivation and the characteristics of the coefficient of correlation will be of interest to the advanced student engaged in statistical research. Such a student will require a much fuller account of the mathematical principles and processes involved than could be undertaken in the present chapter, which is restricted to a simple statement of the formula for determining the coefficient, and to a simple illustration of the method of finding it in the case of two variables.

The method of finding the standard deviation for a series of numbers has been explained in a preceding chapter (see page 107). The coefficient of

[1] See, for example, G. Udny Yule's *An Introduction to the Theory of Statistics,* Chaps. IX, X, XI, XII; and A. L. Bowley's *Elements of Statistics,* Part II, Section VI. For an illustration of the use of the correlation coefficient in testing the validity of certain economic hypotheses, see Prof. Persons' article "The Correlation of Economic Statistics" in the *American Statistical Association's Quarterly,* Dec., 1910. General reference may be made, also, to the *Journal of the Royal Statistical Society* for discussions of methods of measuring correlation.

correlation measures the closeness of correspondence in the fluctuations of two series of numbers with reference to the arithmetic means of the series. It is a fraction whose numerator is the sum of the products of the corresponding deviations, and whose denominator is the product of the standard deviations of the series, and the number of observations or comparisons. If two series, X and Y, be taken, and the deviations from the mean of the two series be represented by $x_1, x_2, x_3 \ldots x_n$, and $y_1, y_2, y_3 \ldots y_n$, respectively, the standard deviations by σ_1 and σ_2, the number of observations by N, and the coefficient of correlation by r, the formula for finding the coefficient is

$$r = \frac{\Sigma xy}{n\sigma_1\sigma_2}$$

The method of finding the coefficient is simply illustrated in the following table, taken with slight modifications, from the article by Professor Persons on "The Correlation of Economic Statistics."[1]

X	Y	x	y	x^2	y^2	xy	
1	2	−2	−4	4	16	8	$M_1 = 3$
2	4	−1	−2	1	4	2	$M_2 = 6$
3	6	0	0	0	0	0	$\sigma_1 = \sqrt{2}$
4	8	+1	+2	1	4	2	$\sigma_2 = \sqrt[4]{2}$
5	10	+2	+4	4	16	8	$r \frac{20}{\sqrt[4]{2}.\sqrt[4]{2}} = 1$

Professor Persons explains that in making up the two series —

. . . . the numbers were chosen so that for an increase of 1 unit in the X series there is an increase of 2 units

[1] *Amer. Statist. Assn. Quarterly*, p. 299, Dec., 1910.

in the Y series. Thus the correlation is perfect and r equals +1. If the Y series had been [10, 8, 6, 4, 2] the X series remaining the same,) the value of r

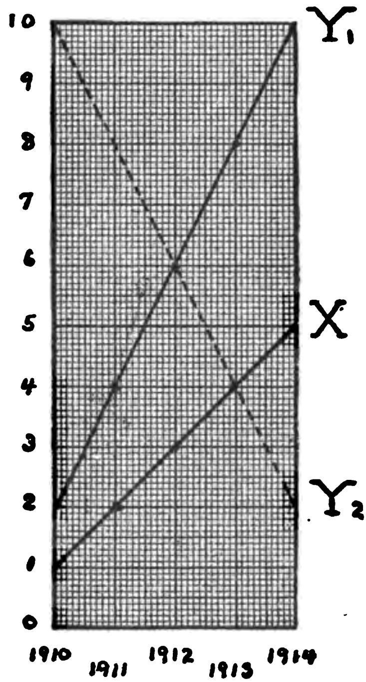


FIG. 1

would have been —1. Thus —1 stands for perfect *negative* correlation, an increase in one series corresponding to a decrease in the other. It should be noted

in this connection that the coefficient of correlation (r) cannot be less than -1 nor more than $+1$.[1]

If it be assumed that the number of the X and Y series relate to specific years, say to the year 1910-1914, they may be represented graphically as shown

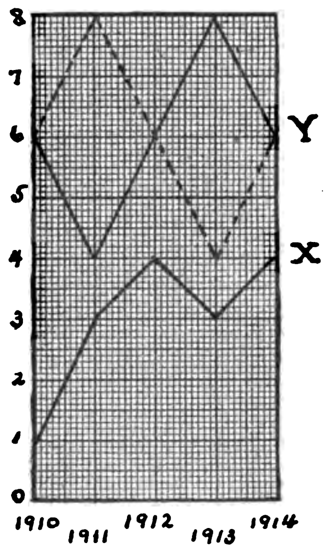


FIG. 2

in Fig. 1, in which the lines giving perfect positive correlation (X and Y_1) and perfect negative correlation (X and Y_2) are plotted.

While the lines plotted on Figure 1 represent spe-

[1] For proof of this statement reference is made to A. L. Bowley, *Elements of Statistics*, p. 319.

cific cases of perfect positive and perfect negative correlation, it may be observed that any other two straight lines imposed upon the diagram would equally well represent perfect correlation either positive or negative, since the absolute divergence or convergence of any two straight lines is constant from year to year.

By reference to the formula it will be apparent that r can equal O only in case the sum of the product deviations equals O. Such a case is illustrated in Figure 2.

$$\begin{array}{ll} X = 1,\ 3,\ 4,\ 3,\ 4 & M_1 = 3 \\ Y_1 = 6,\ 4,\ 6,\ 8,\ 6 & M_2 = 6 \\ Y_2 = 6,\ 8,\ 6,\ 4,\ 6 & M_3 = 6 \\ \qquad\qquad r = O & \end{array}$$

Figure 2, also, illustrates simply specific cases of $r = O$, or no correlation, negative or positive, between X and Y. An infinite number of other cases of no correlation might be arranged by modifying the series arbitrarily.

It will be obvious that each other value of r, besides $+1$, -1 and O, may be represented by various arrangements and compositions of X and Y, indicating that the same degree of correlations may obtain between X and Y in an infinite number of cases or arrangements of the series. Figure 3 illustrates one of the infinite number of cases of imperfect correlation, each of which gives as the coefficient, $r = \pm.98$. The series is as follows:

$$\begin{array}{ll} X = 1,\ 2,\ 3,\ 4,\ 5 & M_1 = 3 \\ Y_1 = 1,\ 2,\ 4,\ 7,\ 11 & M_2 = 5 \\ Y_2 = 11,\ 7,\ 4,\ 2,\ 1 & M_3 = -5 \\ & \ r = \pm.98 \end{array}$$

The number of X, Y arrangements, or compositions representing any given value of r, is infinite, and the number of values of r, between -1 and $+1$ is infinite. As regards correlation, therefore, the coefficient r provides a perfect measure of infinitely fine graduation.

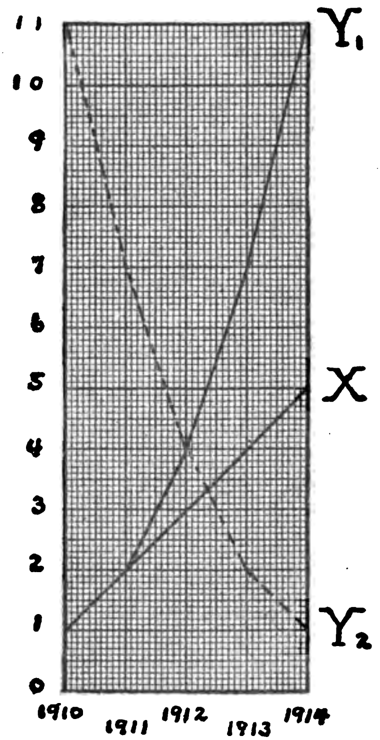


FIG. 3

In practical statistical work, any two series of numbers may be taken as X and Y. Mr. Yule's illustrations of correlation include, for example, biological measurements; ages of husbands and wives; stature of fathers and sons; number of children of mothers and of their daughters; call discount rates, and percentage ratio of reserves to deposits; proportion of male births, and number of births; earnings of agricultural laborers, and percentage of population in receipt of poor relief; percentage of males over sixty-five years of age in receipt of poor relief, and number relieved outdoors to one indoors; infantile mortality per 1,000 births, and general mortality per 1,000 living; persons married per 1,000 population (marriage-rate), and values of exports and imports per capita, each year 1855-1904.

In the last case noted, in order to eliminate the effect of the long-time trend of the marriage-rate, and separate out the annual fluctuations, for comparison the differences between the annual marriage-rates and nine-year averages are taken instead of the annual marriage-rates themselves.

Either the X series or the Y series, or both the X and the Y series may represent means of frequency distributions, the X series, for example, representing the mean age of husbands classified by age of wives, and the Y series mean age of wives classified by age of husbands. If in this case no relationship or correlation obtained between the age of the husband and the age of the wife, the age distribution of husbands would be the same for each

class of husbands defined by age of wife, and similarly of wives classified by age of husbands. The mean age of husbands would not vary with age of wife, nor the mean age of wives with age of husband. In fact, however, the ages of husbands and wives are closely correlated, and this correlation may be measured by taking the means for the husband classes as the X series, and the means of the wife classes as the Y series, finding the deviations from the mean age of husbands and wives and proceeding as has been explained.

BIBLIOGRAPHY

To the student who wishes to pursue further the subject of statistical theory, the following short list of works in English is recommended.

BAILEY, W. B., *Modern Social Conditions.* The Century Company, New York, 1906.

BOWLEY, ARTHUR L., *Elements of Statistics.* 3rd Ed., London, 1908.

BOWLEY, ARTHUR L., *An Elementary Manual of Statistics.* London, 1910.

BRINTON, WILLARD C., *Graphic Methods for Presenting Facts.* The Engineering Magazine Co., New York, 1914.

DAVENPORT, C. B., *Statistical Methods.* John Wiley & Sons, New York, 1904.

ELDERTON, W. PALIN, *Frequency-Curves and Correlation.* London, 1906.

ELDERTON, W. P. and E. M., *Primer of Statistics.* London, 1909.

FARR, W., *Vital Statistics.* London, 1855.

KING, W. I., *The Elements of Statistical Method.* The Macmillan Company, New York, 1912.

NEWSHOLME, A., *The Elements of Vital Statistics.* London, 1892.

YULE, G. UDNY, *An Introduction to the Theory of Statistics.* London, 1911.

ZIZEK, DR. FRANZ, *Statistical Averages.* Translated by W. M. Persons. Henry Holt & Co., New York, 1913.

INDEX

ia/a